

Contents

1	Introduction	1
2	Related work	3
2.1	Datasets of human behavior	3
2.2	Models for behavior understanding	4
3	The EPFL-Smart-Kitchen-30	4
3.1	The EPFL-Smart-Kitchen setup	5
3.2	Data collection procedure	5
3.3	Estimation of 3D motions	5
3.4	Annotation of fine-grained actions and coarse activities	5
4	Multimodal action and motion understanding benchmarks	6
4.1	Lemonade: Language models Evaluation of MOtion aNd Action-Driven Enquiries	6
4.2	Action recognition benchmark	7
4.3	Action segmentation benchmark	8
4.4	Situated full-body motion generation benchmark	9
5	Conclusion, future work and impact	10
A	EPFL-Smart-Kitchen platform	3
A.1	Details on devices and sensors	3
A.2	Synchronization and calibration of the sensors	3
A.3	List of recipes	4
A.3.1	Omelette and Tomato Salad	4
A.3.2	Ratatouille	4
A.3.3	Risotto	5
A.4	Pad Thai	5
A.5	List of kitchen items	5
B	Details on action annotations	5
B.1	Action distributions	5
B.2	Verbs descriptions	5
B.3	Clarification of special Food/Objects	8
B.4	Annotator training	8
B.5	Validation of action annotations	9
C	Automatic 3D body/hand motion estimation	9
C.1	2D keypoint estimation	9
C.2	Identity tracking	9
C.3	Confidence-based 3D keypoint lifting	10
C.4	3D mesh fitting	10
C.5	Assessment of Pose Estimation Quality	10
D	Protocols	11

D.1	Ethics protocol	11
D.2	Data collection protocol	12
E	Compute resources	12
F	Supplementary material for benchmarks	14
F.1	Details on Lemonade	14
F.1.1	Comparison with other benchmarks	14
F.1.2	Question distributions per category	14
F.1.3	Intuition behind categories and subcategories	14
F.1.4	List of questions	15
F.1.5	Question engineering	16
F.2	Visual Language Model (VLM) Evaluation	31
F.3	Egocentric reprojection of the pose data	32
F.4	Details on the Action Recognition Benchmark	32
F.4.1	Additional results	32
F.4.2	Data Preprocessing	32
F.4.3	Hyperparameters	33
F.4.4	Hand-cropped video extraction	33
F.5	Details on the Action Segmentation Benchmark	33
F.5.1	Models	33
F.5.2	Additional results	34
F.6	Fine-grained performance in action segmentation body vs hands	34
F.6.1	Data Preprocessing	34
F.6.2	Hyperparameters	35
F.6.3	Feature engineering details	35
F.7	Details on the Full-body Motion Generation Benchmark	36
F.7.1	Data Preprocessing	36
F.7.2	Evaluator training	36
F.7.3	Tokenizer performance	36
G	Datasheets for datasets	37
G.1	Motivation	37
G.2	Composition	37
G.3	Collection Process	38
G.4	Preprocessing/cleaning/labeling	39
G.5	Uses	40
G.6	Distribution	40
G.7	Maintenance	41

Appendix to the EPFL-Smart-Kitchen-30 dataset and benchmarks

A EPFL-Smart-Kitchen platform

A.1 Details on devices and sensors

We installed nine Kinect Azure cameras [57] inside the kitchen and asked the subjects to wear the HoloLens 2 headset [85] to record video streams. The RGB streams are recorded at 30 FPS, and the depth streams are recorded at 10 FPS. We also estimated the head position, hand poses, and eye gaze data at 30 FPS from the HoloLens 2.

Apart from the video streams, we also installed six wired IMU sensors (Adafruit BNO055 sensor) on some of the appliances (fridge/cupboards) and two wireless IMU sensors (Axiverty AX3 sensor) on the frequently used tools (knife/spatula). The wired IMU sensors were connected to an Arduino, which also recorded the audio signals from the Kinect Azure cameras and then sent them to the IMU sensors for synchronization.

The wireless IMU sensors were synchronized based on the timestamps recorded during data collection. The wired IMU sensors were recorded at 30 FPS, while the wireless IMU sensors were recorded at 100 FPS (Figure A.1 B). The IMU data, especially from the knife and spatula, is captured at a higher temporal resolution than the video recordings. High-frequency data contains rich signals that are important for analyzing complex, fine-grained manipulations through precise event timings or oscillation profiles.

A.2 Synchronization and calibration of the sensors

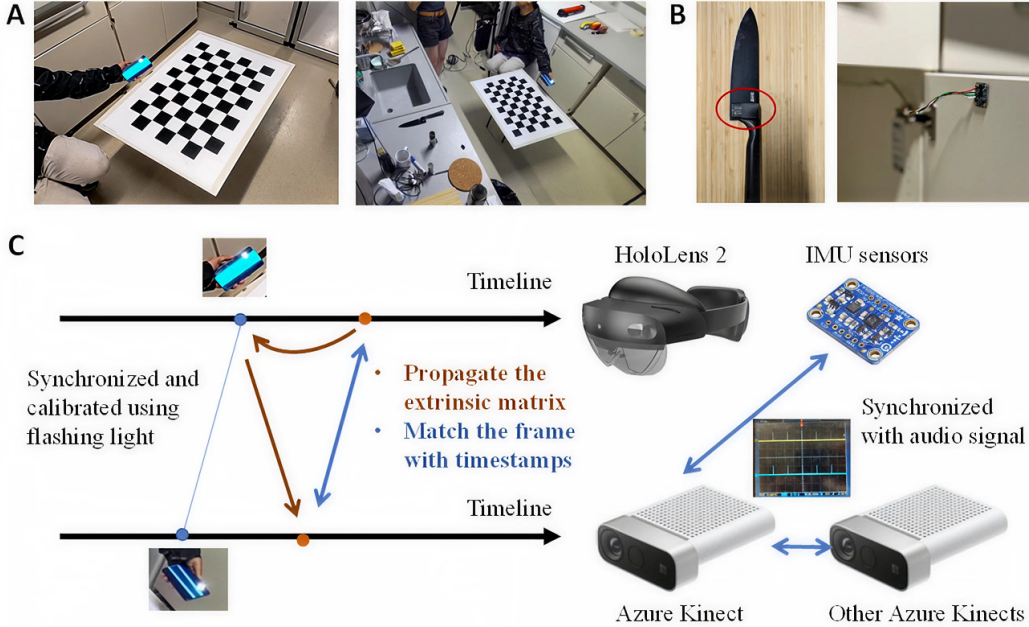


Figure A.1: **Calibration and IMU sensors and synchronization** A). We use an A0-size checkerboard to calibrate egocentric and exocentric cameras at the synchronization timestamp. B). We attach wired IMU sensors to the appliances (fridge/cupboard) inside the kitchen and wireless IMU sensors to the frequently used tools (knife/spatula). C). We synchronize the egocentric and exocentric cameras using the offset of a flashing light and synchronize exocentric cameras and IMU sensors using audio signals.

We calibrated the egocentric and exocentric cameras using an A0-size checkerboard (Figure A.1 A). Since, during a session, the egocentric camera is constantly moving, we calibrate it only at the calibration timestamp. To calibrate the egocentric camera with exocentric cameras for the entire session, we leveraged the head poses obtained from the HoloLens 2 and performed an extrinsic matrix propagation to the calibration timestamp.

As is shown in Figure A.1 C, we synchronize the egocentric and exocentric cameras using the offset of a flashing light at the beginning of the session and use the recorded timestamps to match the captured frames. The Kinect Azure cameras can output regular audio signals; we used these signals to synchronize all the Kinect Azure cameras and the IMU sensors. We validated the calibration and synchronization by projecting the regressed 3D body poses to all the views and manually checking whether the poses were aligned with the body in all the views at randomly selected timestamps.

A.3 List of recipes

Overall, participants are required to cook five times, preparing a different recipe each time, with the exception of ratatouille (section A.3.2), which they prepare twice. A subset of those sessions is included in the EPFL-Smart-Kitchen-30.

A.3.1 Omelette and Tomato Salad

Tomato salad

1. Dice 2 tomatoes
2. Mix one spoon of oil with salt, pepper and balsamic vinegar
3. Pour the dressing on the tomato salad
4. Stir the salad for at least 2 minutes to allow the tomatoes to soak up the dressing

Omelette

1. Beat 3 eggs and season them with salt and pepper
2. Heat the oil in a pan over a medium-low heat
3. Pour the eggs into the pan, tilt the pan ever so slightly from one side to another to allow the eggs to swirl and cover the surface of the pan completely
4. Let the mixture cook for about 20 seconds then scrape a line through the middle with a spatula
5. Tilt the pan again to allow it to fill back up with the runny egg
6. Repeat once or twice more until the egg has just set
7. Fold gently in half with the spatula

A.3.2 Ratatouille

1. Put a casserole of salted water to boil.
2. Heat 2 tablespoons of oil in a pan over medium heat, add the basil stalks and thyme leaves. Cook on a medium heat for 2-3 minutes. During this time, chop the zucchini, eggplant, and mushrooms.
3. When the water boils, add the pasta for 7-10min depending on the type. Then drain them, drizzle them with olive oil, and keep them aside.
4. Then add the zucchini, eggplant, and mushrooms (you may need to do this in batches) and fry until golden and softened (approximately 5 minutes). During this time, deseed and finely chop the pepper.
5. Add the pepper for another 3 minutes. During this time, cut the tomatoes in 4.
6. Stir in the tomatoes, the balsamic, and a good pinch of sea salt and black pepper. Mix well, breaking up the tomatoes with the back of a spoon. Cook for 5- 7 minutes, stirring the vegetables now and again, until reduced, sticky and sweet.
7. Add some basil, finely grate in the lemon zest, and adjust the seasoning if needed.

A.3.3 Risotto

1. Bring 1.5L of water to a boil in a saucepan and pour in a vegetable stock cube, stir. As soon as it boils, reduce the heat to low and let the broth simmer to keep it warm.
2. Melt the butter in a second pan over medium heat. Add 250 g of risotto rice and cook for about 3 minutes, stirring well, until it becomes translucent. Pour in 10 cl of white wine and cook, stirring frequently, until the rice completely absorbs it.
3. Pour a ladleful of vegetable stock into the pan and continue cooking, stirring occasionally, until the stock has completely evaporated. Once it has evaporated, add another ladleful of broth and wait until it is absorbed again. Repeat several times, adding ladles of broth to the risotto as you go, until the risotto is cooked.
4. Meanwhile, make a vegetable salad. Cut 1-half of cucumber into wedges. Cut the radish into slices, and cut the surimi into slices. Cut the avocado in 2 halves, remove the core, and cut the avocado into thin slices. Peel and grate the carrot and mix everything in a bowl.
5. Prepare the dressing by mixing the soy sauce and the sesame oil. Pour the dressing over the vegetables and toss for at least 1 minute. Sprinkle with sesame seeds.
6. When the risotto is cooked, add a spoonful of butter and 3 spoonfuls of grated Parmesan cheese. Add salt and pepper to taste and stir well.

A.4 Pad Thai

1. Dip the noodles in hot water to soften them
2. Dissolve a portion of tamarind paste in 100 ml of hot water, stir, and filter the mix to obtain tamarind juice.
3. Cut the tofu in cubes, the shallot, and the onion in slices.
4. Heat up the frying oil in a pan and cook the tofu until it becomes golden, reserve for later
5. Add the drained noodles and cook them with the juice in the pan, add water if the noodles absorb too much.
6. Add the egg to the pan next to the noodles and cook the egg before mixing the noodles, you should obtain separate white and yellow parts.
7. Add the tofu, the peanuts, and the dried shrimp
8. Remove from the heat and add the soy sprouts, the onion, and the shallot, serve with a lemon slice.

A.5 List of kitchen items

Table A.1 lists all items present and used in the kitchen by the participants.

B Details on action annotations

B.1 Action distributions

Our EPFL-Smart-Kitchen-30 provides action segments of various lengths with a long-tail distribution. To support this, we show the duration of verbs and nouns in Figure B.1.

B.2 Verbs descriptions

Each verb is provided with a detailed description to ensure specific usage and avoid confusion during annotations. Below is an exhaustive list of the verb descriptions along with general annotation rules. We note that these instructions could be used for NLP-based systems in the future.

- *Open/Close*: Opening/Closing objects such as Pot, Package, Butter, Tamarind paste, Water, Oil, Sauce, or appliances/objects such as Fridge, Cupboard, Drawer.
- *Cut*: Operation of Cut an object or food when Carrying a Knife. Exceptions: The action of opening a package with a Knife should be defined as Open.

Table A.1: Exhaustive list of items present in the EPFL-Smart-Kitchen

<i>Appliances</i>					
Fridge	Counter	2×Stoves	Ventilation	Sink	Drawers
Cupboards	Drying rack				
<i>Objects</i>					
Spatula	Knife	Bottle	3×Pot	Pot lid	Peeler
Recipe	Sponge	2 Salad bowls	Cup	Cutting board	Bowl
Colander	Doser glass	Grater	Tissue	Brush	Cleaning gloves
Paper plates	Spoon	Sink Sprayer	Soap	Pasta Spoon	Tissue
Whip	3×Pan	Towel	Trivet	Fork	
<i>Food</i>					
Onions	Tomatoes	Avocado	Bell Pepper	Radish	Zucchini
Cucumber	Mushrooms	Shallots	Carrots	Eggs	Butter
Surimi	Shrimps	Poultry Broth	Pasta	Noodles	Rice
Vegetable Broth	Olive oil	Frying oil	White wine	Balsamic vinegar	Fish sauce
Basil	Paprika	Tamarind paste	Nutmeg	White sesame	Parsley
Salt	Pepper	Water	Eggplant	Soy sprouts	Thyme
Sesame oil	Lemon	Tofu			

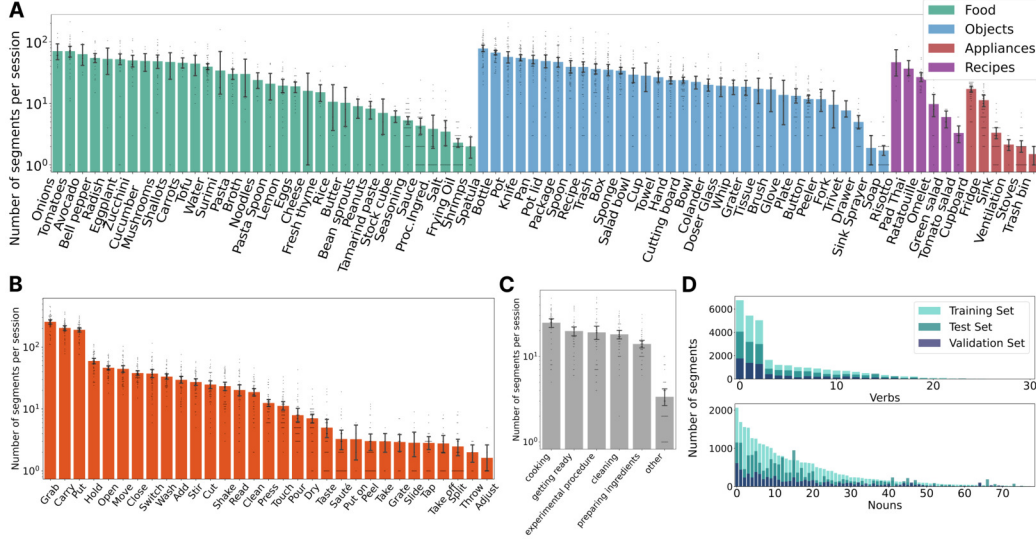


Figure B.1: **Number of action segments.** (A) Nouns, (B) Verbs, (C) Activity (D) Number of segments per split for nouns (top) and verbs (bottom). N = 49.

- *Wash*: Washing an object or food requires the use of a certain amount of Water and Soap and should happen in the sink area. To be contrasted with Clean which is reserved for areas apart from the Sink.
- *Clean*: Cleaning an object/food requires the use of a Tissue, Towel, sponge, or Hand. It can happen in any area apart from the Sink area. It could involve motions such as wiping or removing food from the Knife in a Cut action.
- *Stir*: Motion of stirring which can happen in Salad Bowls, Pan, Pot, etc.
- *Wait*: This action is not related to any noun. Wait action should last for at least 3 seconds, otherwise there is no need to label this segment.
- *Put*: Action of setting down an object.
- *Split*: Action of separating Processed Ingredients in two parts. To be contrasted with Cut which mostly happens with which is also splitting but with a Knife. Cut must be prioritized over Split if in hesitation.

- *Touch*: The action of barely touching and getting in contact with an object and having absolutely no consequences, can be involuntary. To be contrasted with Tap (repetitive movement) and Press (voluntary and has a certain degree of consequences), with more force). Touch does not need to be annotated during other behaviors that involve touch (e.g. Stir, Tap, ..)
- *Peel*: Removing the skin of a specific food. Can happen with a Peeler, a Knife, or directly with Hands.
- *Pour*: The action of setting down transferring a liquid (Water, Oil, Sauce, Broth) from one recipient into another recipient (such as from a Bottle into a Pan, or from a Bowl into the Sink). Note that Tamarind paste is also considered as liquid and follows the Pour rule.
- *Sauté*: Sautéing represents the movement where the cook shakes the Pan instead of Stir. It is only used together with Pan.
- *Taste*: The action of taking putting food in the mouth. The action should be annotated from the moment the food reaches the mouth of the participant and should be following Grab and Carry.
- *Tap*: Tap with the Spatula, Spoon, or Knife often to remove the stuck residues.
- *Add*: Add a set of ingredients to another set of ingredients. Add has a higher priority than Put. For example, we prefer Grab/Carry/Add Pasta instead of Grab/Carry/Put Pasta when the subject wants to cook the pasta and add it to the Pan.
- *Throw*: Defined when the object is thrown in the air.
- *Carry*: Defined when the interacting object is lifted. Intermediary action defined in between Grab and Put/Add.
- *Hold*: Defined when the person tries to stabilize the object with almost no placement. Hold must be annotated also during Cut if the participant Holds with the contralateral hand.
- *Move*: Move is an action that involves the displacement of an object without lifting it (slide). Move is not followed by Put.
- *Dry*: Action happening after Wash. Drying requires the use of a Towel or Tissue. Involve the action of wiping on an object. To be contrasted with Shake when performed without Towel or Tissue.
- *Grab*: Grab an object, must typically be followed by Carry (lifted) and eventually Put or Hold (not lifted). Grab is defined from the moment when the reaching movement starts and ends at the hand-object contact instant. Note that some sequences can also have Grab followed by Hold then followed by Carry.
- *Read*: Applies to Recipe but also to any object with associated text such as Package or Bottle.
- *Press*: Typically Press Button when the participant sets the Stoves or Press Hands when the participant Open/Close the HoloLens menu by pressing their wrist. To be contrasted with Tap (repetitive movement with an object) and Touch (No consequences of the action). Press Button can be defined once for repetitive instances of the actions as long as the arm of the participant does not retract but must not intersect with Slide Button (For the slide button in the middle of the stoves)
- *Slide*: Correspond to the sliding of a hand or an object on a surface or other object. The participants have the possibility to Slide the Button of the stoves.
- *Shake*: Move an object from side to side with a forceful, jerky movement. Should be used instead of Dry when relevant.
- *Squat*: Knees bent and one's heels close to or touching one's buttocks or the back of one's thighs. The Squat action is not associated with any noun. Often happens in front of the fridge. It represents the whole state rather than only the movement from the start of the crouching to the end of the standing-up action.
- *Switch*: Switch an object from one hand to another, annotate when both hands are in contact with the object.
- *Grate*: It happens when the subject uses the Grater (object) to process the food (e.g., carrots or cheese).

- *Put on/ Take off*: They can be used exclusively with Glove or other clothes
- *Take(with tool)*: to describe the action when the object is lifted by a tool. It is used to describe a status, is not necessarily followed by Put.

B.3 Clarification of special Food/Objects

- Tamarind Paste is only used in the Pad Thaï recipe as a paste mixed with water. After dilution please refer to it as Sauce
- Broth is only used in the Risotto recipe after the stock cube is Mixed with the water, the participant typically Add Broth to the Rice
- To avoid confusion, Frying Oil should only be used when Frying Oil is added to the pan. Otherwise, it is considered a Sauce.
- *Sauce*: Any cooking liquid including oils, vinegar, wine and liquid mixes. Adding seasoning to a sauce does change its nature. Exception: Frying oil
- *Salt*: To avoid confusion, Salt must be only used when adding salt to the pasta water. Otherwise it is considered as Seasoning.
- Pasta is only present in the ratatouille (wheat pasta)
- Noodles are used in Pad Thaï (rice noodles).
- Seasoning involves any herbs, salts or spices added with a shaking motion. Exception: Salt when added to the water of the pasta.
- *Bowl and salad bowl*: The EPFL-Smart-Kitchen possesses 2 metallic salad bowls of different sizes and metallic. Any other spherical-shaped container can be considered as a Bowl. It can involve the small yellow bowl and eventually the lunch boxes.
- Processed ingredients are normally referred to as “uncooked food”. If the food mix is cooked, then try to call it with the corresponding recipe.
- Specific rules for annotation using a limited set of words:
- Actions can occur simultaneously
- Food is defined by their specific names only if they are targeted by the action.
- When food is processed but not yet cooked please use processed ingredients. By definition, this does not apply to Salads (as it does not require cooking).
- If all ingredients are added to the mix and the participant manipulates the mix, please use the name of the corresponding recipe instead. For instance, in the Ratatouille recipe, when the tomatoes, zucchini, and eggplant are mixed together, then it is labeled as ratatouille.
- An object should be defined as Trash as late as possible before going to the Trash bin. The object should keep its name (if known) as long as possible, the last action should still include Trash in its name. (example Grab Zucchini, Put Trash)
- Please be as precise as possible on the object interaction, for instance, the participant will never Carry a seasoning but a bottle - they Carry a Bottle. However, they may add the Seasoning.
- All the actions should last for at least 5 frames.

B.4 Annotator training

Our research utilized the services of a third-party commercial annotation company specialized in data labeling.

To ensure high-quality labels, we trained annotators with two mini-batches of data. Specifically, we first asked the annotators to review our annotation requirements (see section B.2 and section B.3). After that, we gave them the mini-batch of data for initial annotation. Meanwhile, two authors of this paper manually annotated the mini-batch of data. The annotated actions were merged by manual check to serve as the gold standard, which was given to the annotators after they finished their own annotations. After two rounds of training, and when annotations from all the annotators could achieve an F1 classification score larger than 0.9, the annotators started to collect ground truth annotations.

B.5 Validation of action annotations

We perform two ways of annotation validation jointly to ensure the annotation quality. Firstly, we asked additional annotators to randomly check the annotations of certain clips. Videos with unsatisfactory clip annotations will be returned to the annotators for second-round annotation. Secondly, we applied a rule-based check for all the annotations. For example, we check if the ‘Cut’ action segments are inside the ‘Carry Knife’ action segments. Unsatisfactory action segments were returned to the annotators for second-round annotation as well.

C Automatic 3D body/hand motion estimation

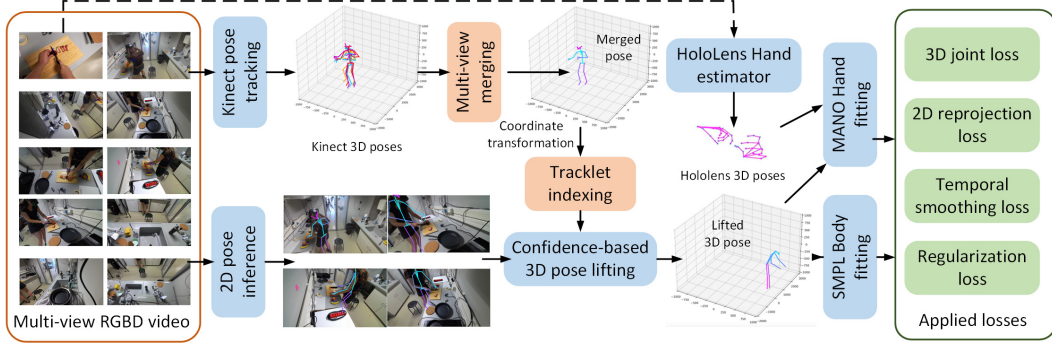


Figure C.1: **Body/Hand motion estimation pipeline.** We effectively use multi-view and multi-modal information to regress accurate 3D body and hand meshes across the whole video sequence.

To efficiently and accurately capture a large volume of hand and body motions from the captured videos, we propose a body/hand motion estimation method that can effectively use multi-view and multi-modal information (Figure C.1). Specifically, the body/hand motion estimation method can be divided into four steps: 2D keypoint estimation, identity tracking, 3D keypoint lifting, and 3D mesh fitting. We will detail each step in the following sections. Furthermore, we also compare the bone length variance between our estimated hand poses with the HoloLens 2 estimated hand poses to show the effectiveness of our estimated poses.

C.1 2D keypoint estimation

To balance performance and efficiency, we choose to use RTMPose [31] for 2D pose estimation; the model is available in DeepLabCut v3 [53]. Specifically, we use the RTMPose-x model trained on eight pose estimation datasets to extract body poses [31]. Due to the small size of the hand regions in the video, the existing hand detectors often struggle with hand-missing detection and, due to occlusion or interactions, are prone to errors. We chose to use the whole-body model (RTMW-x [31]) trained on 14 public datasets to extract hand poses, because it relies on semantic relations to effectively find the hand regions.

C.2 Identity tracking

The experimentalists occasionally appear in the data collection area to assist the subjects. When this occurs, RTMPose [31] can misalign detection identities for the detections in previous frames. Therefore, before grouping the 2D pose estimation from each camera for 3D pose estimation, it is necessary to identify the 2D pose of the subject in each view across the entire video. To achieve this, we rely on the Kinect body tracking SDK [58] to determine the correct identity of the subject. It estimates the 3D body pose and the tracking identity using the depth images from each camera view. We first transform the 3D body poses into the world coordinate system and merge them to obtain a robust 3D body pose estimation. After that, we consider the dominant identity detected in the video as the subject and project that 3D body pose into each camera view. We use the projected 2D body pose to identify the correct 2D poses estimated by RTMPose [31]. We do not use the merged 3D body pose intermediate results as the final 3D pose results as they are quite jittery.

C.3 Confidence-based 3D keypoint lifting

With identity information, we group the 2D poses of the subject from all camera views for 3D pose lifting. We perform singular value decomposition (SVD) using both the 2D pose results and the corresponding confidence scores from all camera views. Pose joints with scores lower than a certain threshold (0.1) are excluded from subsequent 3D mesh fitting.

C.4 3D mesh fitting

We adapted the multi-view mesh fitting implementation of EasyMocap [1] to estimate the body and hand meshes of our data. Specifically, we utilized lifted 3D poses and 2D poses estimated by RTMPose [31] to fit the SMPL parametric model [62]. We modified the original joint regressor matrix to adapt it to the COCO [44] joint definition. 3D joint loss, 2D reprojection loss, temporal smoothing loss, and regularization loss are minimized to optimize the pose, shape, and global transformation parameters of SMPL [62]. Meanwhile, the 3D hand poses obtained by the HoloLens 2 hand tracking toolkit [59] are estimated from the egocentric view depth image, which contains hand contact information that RGB views may lack. Therefore, we also use the relative 3D hand poses from HoloLens 2 to minimize an additional 3D joint loss for fitting the MANO parametric model [68]. Finally, we use the fitted SMPL and MANO models to obtain the final 3D body and hand poses (Figure C.2). We discard frames where the fitted pose significantly differs in Euclidean distance from the lifted pose, as this discrepancy is likely caused by artifacts in the fitting algorithm.

C.5 Assessment of Pose Estimation Quality

Creating 3D ground truth is notoriously difficult. We measure the quality of the 3D pose estimation pipeline by manually annotating 4,947 frames with 2D keypoints for a fixed, random selection of 14 keypoints from all exocentric views. The frames were extracted using DeepLabCut’s frame extraction pipeline; specifically, we sampled distinct visual frames via k-means [53].

We then compare the mean euclidean distance between our predicted poses and the triangulated annotated ground truth data. Our pose estimation shows an average error of $52.75 \text{ mm} \pm 53.26 \text{ mm}$, where $31.73 \text{ mm} \pm 43.06 \text{ mm}$ is from the hands and $57.69 \text{ mm} \pm 54.22 \text{ mm}$ is from the body. These findings are based on 469 triangulated poses (Figure C.3). Our pose estimation provides results closer to the ground truth compared to the original merged 3D body pose estimations from Kinect Azure cameras and HoloLens 2 3D hand poses (Figure C.3). The Kinect Azure 3D pose estimation has an error of $118.37 \text{ mm} \pm 177.76 \text{ mm}$ (for the body), while the HoloLens 2 estimation has an error of $86.20 \text{ mm} \pm 297.28 \text{ mm}$, comparing only frames with visible hands in an egocentric view.

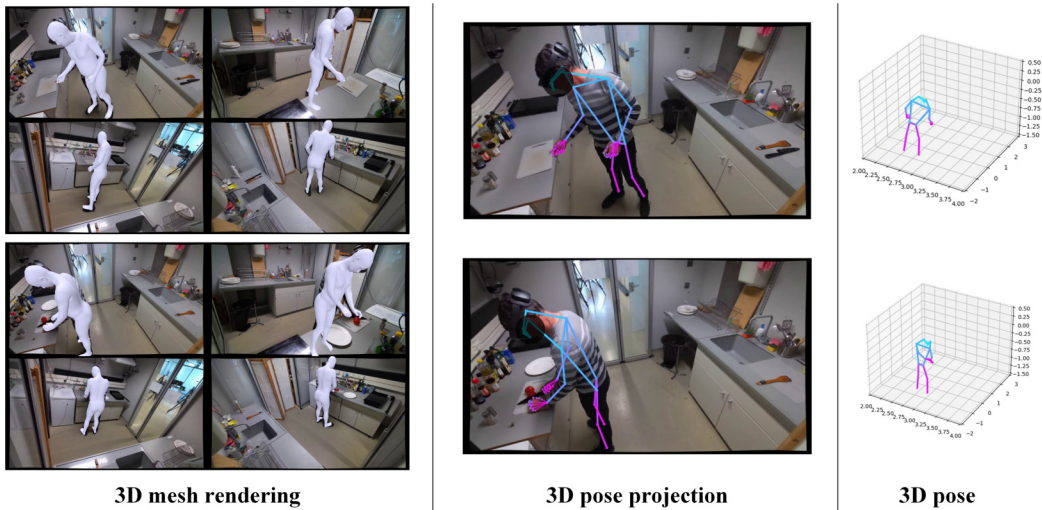


Figure C.2: **3D pose examples.** Examples of the estimated Body/Hand meshes (left), reprojected 3D pose (center) and 3D pose (right).

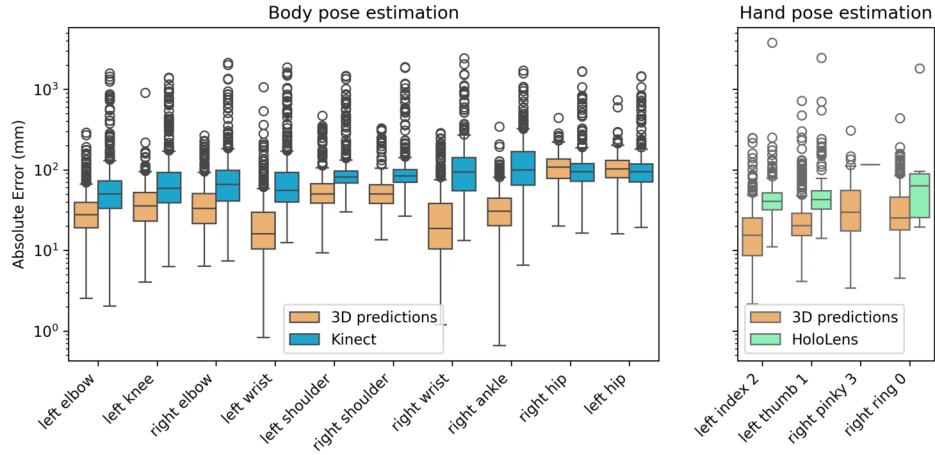


Figure C.3: Absolute error with triangulated ground truth comparing our 3D pose estimation (N=469) to 3D body prediction from the Kinect Azure cameras (N=469) and HoloLens 2 hand pose estimation (N=157). 3D hand pose estimation

D Protocols

D.1 Ethics protocol

The study was approved by the Swiss Ethics Committees on research involving humans (Project ID: 2022-00493). Following the approved protocol, the participants sign a research agreement before they start participating. Here, we share a summary for the community (All content in *italics* is copied from the original ethics protocol):

- **Motivation of this study :**

The goal of this project is to build a database of subjects using a kitchen in a “natural world” scenario. We are collecting data on healthy subjects and on subjects with motor deficits. These data will help us to build new metrics (standardized evaluations) using machine learning to monitor patient’s progress during their rehabilitation process or during the use of assistance devices.

- **Procedure of a typical cooking session :**

If you accept to participate in our project, you will use a kitchen equipped with wall-fixed cameras, one head-fixed camera that will record your arms and track your eyes, and cooking utensils equipped with movement and force sensors. You will also answer some surveys. You will be allowed to use the kitchen at your own convenience, taking into consideration the following conditions:

- (1) *Preparation of the meal*
- (2) *Cooking (for instance cooking pasta) and preparation of drinks (for instance coffee, tea).*
- (3) *Cleaning of the cooking utensils for the next participant.*

- **Benefits and risks or constraints related to the participation in this study :**

By participating in this study, you will only be exposed to minor inconveniences, such as the recording of your movements and carrying the head-mounted camera. There are no particular risks related to the recording except the two cited in a normal kitchen (hot water, use of a knife, etc . . .)

- **Anonymization of the collected data :**

If your face or any distinctive feature (such as tattoos) appears on one of the videos, they will automatically be detected and blurred at the end of the recording session. We will only conserve the videos the edited videos with your face, or any other distinctive sign blurred, and delete the original data.

Your data will be accessible to a limited number of researchers working on this study. These data will be pseudoanonymized (which means that your identification information (name, birthday) will be replaced by a code). Your identity key will be secured.

Your data will be useful for the creation of the dataset. At this point, your data will be anonymized (which means that your personal data will be aggregated and protected, preventing anyone from identifying you. Your name will never appear on the internet or on a publication.

- **Optional nature of participation and obligations** *Your participation is completely free. If you choose not to participate or if you choose to participate and reconsider your decision during the course of the project, you will not have to justify yourself. This decision will not have a negative impact on the rest of your medical care. If you choose to participate in this research project, you will be required to follow the instructions and fulfill the requirements of the research protocol.*

D.2 Data collection protocol

The data collection process of EPFL-Smart-Kitchen-30 was reviewed and approved by the Swiss Ethics Committee. Data capture occurs only in the EPFL-Smart-Kitchen at Campus Biotech, Geneva CH. All subjects underwent a MoCA [61] test before starting the first cooking session to assess for mild cognitive impairment; subjects with a score lower than 25 were excluded from the study.

Each subject participated in five cooking sessions ranging from 30 minutes to 1 hour throughout the study. They were presented with four recipes in the following order: Omelet and Tomato salad, Ratatouille, Risotto and green salad, Ratatouille, Pad Thai (see A.3 for the content of the recipes). Ratatouille was cooked twice to study subjects' adaptation. At the start of each session, the participant followed the calibration protocol of the HoloLens 2 headset for eye tracking.

At the beginning of each session, subjects were given the following instructions:

- Prioritize using the knife and spatula because they are equipped with some of the IMU sensors.
- Keep their hand movements within the visual field of the HoloLens 2 camera.

The session started after the calibration and synchronization frames of all 10 cameras were recorded. During each session, subjects could ask for clarifications on the kitchen appliances (particularly the stoves) and cooking methods.

As stated in the ethical protocol, we blurred the participants' faces for all the shared data.

E Compute resources

Table E.1 outlines the details of the computer resources used to collect the video data, estimate the 3D body/hand motions, and train each benchmark model.

Table E.1: Details on computational resources used for data collection, 3D pose estimation and benchmarking.

Data collection	
Computers	We used four computers during data collection to avoid frame losses due to buffer overload. Three computers collected data from three Kinect Azure cameras while the last computer collected exclusively the HoloLens 2 videos.
RAM	$3 \times 32\text{GB}$
3D body/hand motion estimation	
GPUs	$1 \times \text{RTX 3080 Ti}$
RAM	$1 \times 32\text{ GB}$
Question-answering benchmark	
GPUs	$4 \times \text{A100}$
RAM	$1 \times 64\text{ GB}$
Action recognition benchmark	
GPUs	$2 \times \text{A100}$
RAM	$1 \times 64\text{ GB}$
Action segmentation benchmark	
GPUs	GeForce RTX 3090 and A100
RAM	To run experiments in parallel we used computers with different GPU settings to train models on the action segmentation task 64GB

F Supplementary material for benchmarks

F.1 Details on Lemonade

Lemonade is fully available on Hugging Face at <https://huggingface.co/datasets/amathislab/LEMONADE>.

F.1.1 Comparison with other benchmarks

Lemonade is well situated compared to other benchmarks for assessing human motion. Lemonade has the particularity of using ground truth action annotations and pose estimation for designing QA pairs (Table F.1).

Table F.1: Comparison with other QA benchmarks related to motion and movement.

Question answering video datasets	# Videos	# Closed -ended QA	Ego -centric	From action annotations	From pose estimation
ActivityNet-QA [96]	5,800	58,000	✗	✓	✗
MVBench [38]	4,000	4,000	✗	✗	✗
TVBench [13]	2,525	2,525	✗	✗	✗
MLVU [105]	1,730	3,102	✓	✗	✗
MotionBench [29]	5,385	8,052	✗	✗	✗
EgoSchema [51]	5,031	5,031	✓	✗	✗
EgoTaskQA [30]	2,315	40,322	✓	✓	✗
Lemonade (ours)	36,521	36,521	✓	✓	✓

F.1.2 Question distributions per category

Questions are selected and filtered from a pool of 5,116,194 questions, automatically generated through a combination of clips, question types, and answer ranges. The dataset was sampled and filtered based on frame visibility, presence of hands in specific questions, and manual curation. Lemonade includes 18,857 questions in the Behavior Understanding category, with 9,518 in the Perception subcategory and 9,339 in the Reasoning subcategory. There are 8,201 questions in the Long-term Understanding category, with 2,065 in the Session Properties subcategory and 6,136 in the Summarization subcategory. Lastly, 9,463 questions belong to the Motion and Biomechanics category, with 5,916 in the Kinematics subcategory and 3,547 in the Physical Attribute subcategory. Table F.2 shows an exhaustive list of the question types.

F.1.3 Intuition behind categories and subcategories

Behavior understanding questions centered around annotated actions and activities.

- **Perception:** Questions directly asking which behavior are visible in the clip. This loosely compares to traditional tasks on action recognition or action segmentation.
- **Reasoning:** Questions asking about behaviors not visible in the frames (performed before or after). Context information should feed a reasoning method to answer these questions.

Long-term understanding questions use long-duration clips as input.

- **Session properties:** Questions about session duration or participants age. These questions evaluate if general information can be inferred from long clips.
- **Summarization:** Questions on action sequences or number of instances given actions are performed in a session.

Motion and Biomechanics questions take pose estimation as reference information.

- **Physical attributes:** Question related to static pose information.
- **Kinematics:** Question related to speed and motion.

F.1.4 List of questions

Table F.2: Base questions with their identifiers and number of samples. In italic, elements subject to change.

QID	Question examples	# Questions
Behavior understanding — Perception		
0	"What action am I doing ?"	1003
1	"What activity am I doing ?"	1025
2	"How many actions am I doing ?"	1066
3	"What am I doing with the <i>knife</i> ?"	1017
4	"For how long am I <i>holding</i> the <i>pan</i> ?"	987
5	"For how long am I <i>cleaning</i> ?"	1054
13	"What am I <i>pouring</i> ?"	1027
14	"How much time passes from <i>cutting</i> the <i>zucchini</i> to <i>grabbing</i> the <i>zucchini</i> ?"	1037
15	"At what moment does <i>carrying</i> the <i>tomatoes</i> <i>starts</i> in the clip?"	1031
Behavior understanding — Reasoning		
6	"I am currently <i>carrying</i> the <i>tomatoes</i> , what will be the next action(s) ?"	1030
7	"I am currently <i>holding</i> , what was the <i>previous action</i> ?"	1021
8	"I am currently <i>cooking at the stoves</i> , what will my next activity be ?"	1234
9	"What were my previous 3 actions ?"	2096
10	"What will be my next 2 actions ?"	2184
11	"I am currently <i>grabbing</i> the <i>zucchini</i> , what was my <i>previous activity</i> ?"	1008
12	"I am currently <i>pouring</i> , what is my <i>current activity</i> ?"	1037
Long-term understanding — Sessions properties		
16	"What is my age ?"	1025
17	"What recipe am I cooking ?"	1040
Long-term understanding — Summarization		
18	How many times am I <i>grabbing</i> the peeler in this session ?"	1000
19	"How many times am I <i>tasting</i> in this session ?"	988
20	"For how long am I cooking ?"	1004
21	"What is the correct sequence of action ?"	2150
22	"What was the longest action in this session ?"	994
Kinematic and Biomechanics — Physical attributes		
23	"What is the average height of my eyes in this clip ?"	1177
24	"What is the average shape of my <i>right</i> hand in this clip ?"	1200
25	"What is my average trunk bending angle in the clip?"	1170
Kinematic and Biomechanics — Kinematics		
26	"The clip lasts 1.43s, What is the average speed of my <i>left hand</i> in this clip ?"	1187
27	"The clip lasts 1.73s, At what speed am I reaching for the <i>box</i> ?"	1180
28	"The clip lasts 1.6s, At what speed am I putting the <i>bowl</i> down?"	1186
29	"What is the <i>minimum</i> angle between my <i>left shoulder</i> , my <i>right shoulder</i> and my <i>right elbow</i> in this clip ?"	1163
30	"What is the <i>maximum</i> distance between my hands?"	1200

F.1.5 Question engineering

Q0: What action am I doing ?

Rational. This question challenges VLMs to perceive the action present in the clip.

Design. Action segments were sampled and used as reference clips. Answers are actions (verb + noun) or verbs randomly sampled from the action or verb list.



What action am I doing ?

- A: grabbing the trash
- B: adding the cheese
- C: washing the hand
- D: moving the surimi

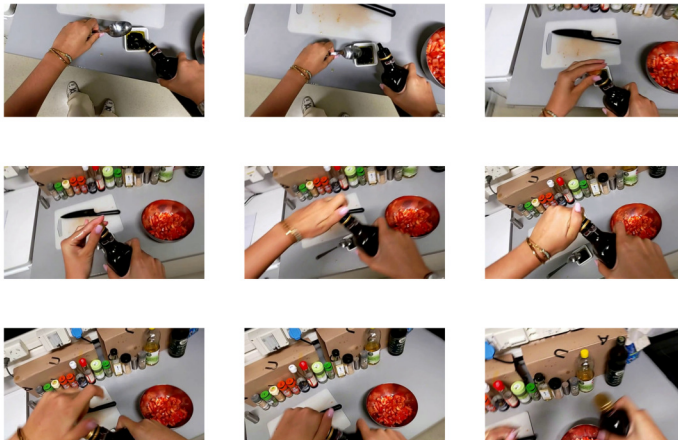
Correct: B

Figure F.1: QID: 0

Q1: What activity am I doing ?

Rational. This question challenges VLMs to perceive the activity present in the clip (coarse-grained action)

Design. Activity segments were sampled and used as reference clips. Answers are actions (verb + noun) or verbs randomly sampled from the action or verb list.



What activity am I doing ?

- A: Gathering supplies
- B: Preparing ingredients
- C: Cleaning up
- D: Setting up and not cooking

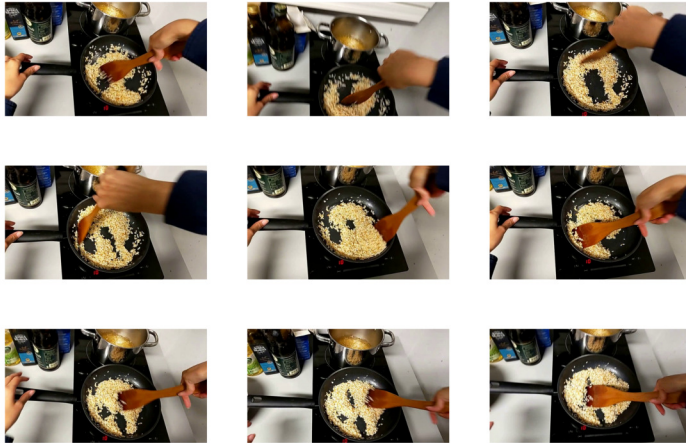
Correct: B

Figure F.2: QID: 1

Q2: How many actions am I doing ?

Rational. Based on the precise definition we give of actions, the number of actions present in a clip is a fixed number. This question challenges VLMs to count the number of actions that happen in a short clip.

Design. For clips ranging from Xs to Ys, we count the number of segments that start within the clip. Answers are randomly sampled from 0 to 5.



How many actions am I doing ?

- A: 3
- B: 4
- C: 5
- D: 1

Correct: A

Figure F.3: QID: 2

Q3: What am I doing with the [NOUN]?

Rational. Prediction of verbs based on nouns.

Design. Action segments were sampled and used as reference clips. We extract the noun from the action for the question and the verb for the answers. Answers are verbs randomly sampled from the verb list.



What am I doing with the cucumber ?

- A: Grab
- B: Split
- C: Pour
- D: Carry

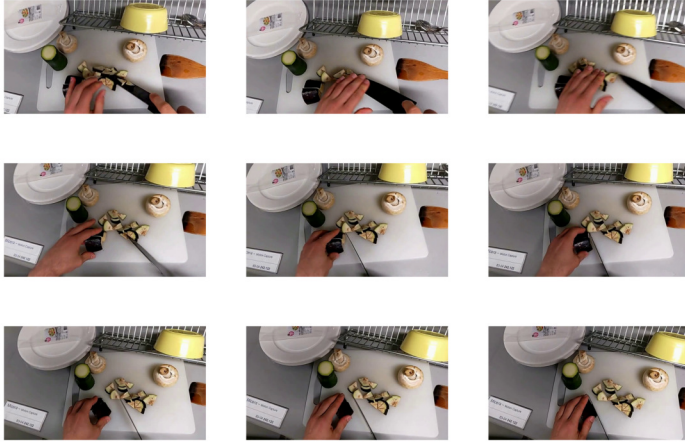
Correct: A

Figure F.4: QID: 3

Q4: For how long am I [VERB] the [NOUN] ?

Rational. Estimating timings from context is difficult both for VLMs and humans. This question challenges models to estimate timings based on context.

Design. Action segments were sampled and used as a reference clip. We extract the correct answer from the clip length. Other answers are samples from windows with sizes corresponding to 90%, 50%, and 20% difference with the correct answers for difficulties in easy, medium, hard respectively.



For how long am I holding the eggplant ?

- A: 2.067s
- B: 2.667s
- C: 3.267s
- D: 3.9s

Correct: A

Figure F.5: QID: 4

Q5: For how long am I [VERB] ?

Rational. Similar to Q4 but based on verbs only.

Design. Similar to Q4 but using clips based on verb segments.



For how long am I carrying ?

- A: 1.333s
- B: 2.333s
- C: 3.333s
- D: 4.333s

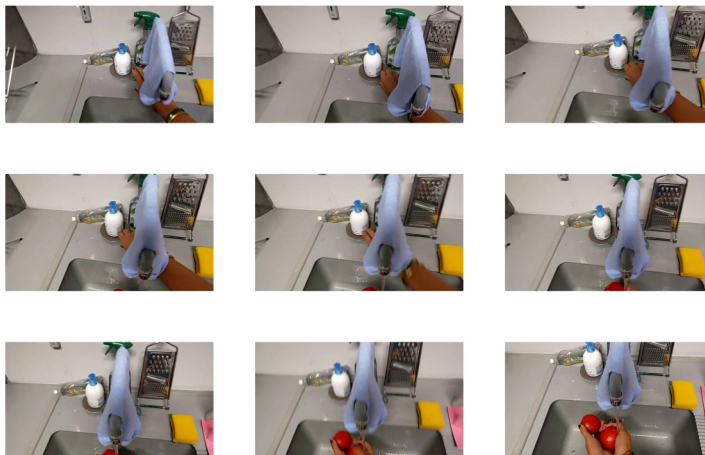
Correct: C

Figure F.6: QID: 5

Q6: I am currently [VERB] the [NOUN], what [BE|past,future] [NEXT/PREVIOUS] action ?

Rational. This question challenges models to reason about actions that would happen before and after the visible clip. This requires a strong understanding of the context and the participants' behaviors. Note that predicting previous actions is more difficult both for humans and models.

Design. Action segments were sampled and used as a reference. The correct answer corresponds to the action that starts after/before the start of the current action. Other answers are randomly sampled from the action list.



I am currently carrying the tomatoes, what will be the next action(s) ?
A: adding the processed ingredients
B: cleaning the pan
C: putting the towel
D: opening the water

Correct: D

Figure F.7: QID: 6

Q7: I am currently [VERB], what [BE|past,future] [NEXT/PREVIOUS] action ?

Rational. Similar to Q6 but based on verbs only.

Design. Similar to Q6 but using clip and answers based on verbs.



I am currently holding, what was the previous action ?
A: grabbing
B: adjusting
C: reading
D: stirring


Correct: A

Figure F.8: QID: 7

Q8: I am currently [ACTIVITY], what [BE|past,future] [NEXT/PREVIOUS] activity?

Rational. Similar to Q6 and Q7 but based on activities only.

Design. Similar to Q6 and Q7 but using clip and answers based on activity.



I am currently cooking at the stoves, what was my previous activity ?

- A: Cleaning up
- B: Other activity
- C: Gathering supplies
- D: Preparing ingredients

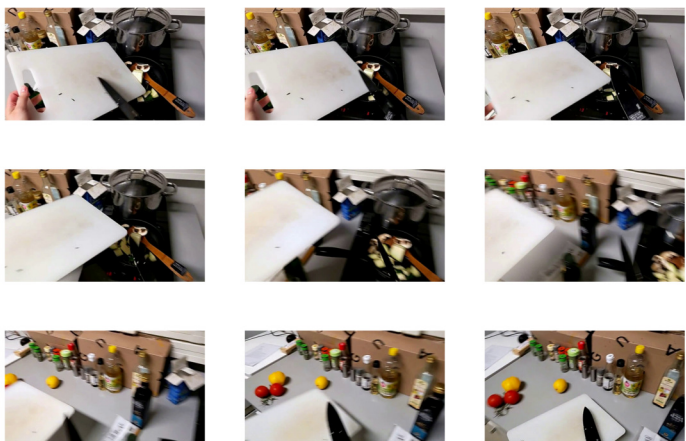
Correct: A

Figure F.9: QID: 8

Q9: What were my previous [NUMBER] actions ?

Rational. The rationale is similar to Q5, Q6, and Q7, with an additional difficulty being the possible answers also sampled from previous actions. The models are required to rank the probabilities that a set of actions is happening right before rather than categorizing them.

Design. The window size is fixed at 50 frames. We extract 2,3,4 (easy, medium, hard) actions from the window before the current clip. Answers prioritize the actions that are present in the video but happening before the extracted actions. The other actions are randomly sampled from the action list.



What were my previous 3 actions ?

- A: closing the cupboard + moving the cutting board + carrying the knife
- B: stirring the mushrooms + carrying the salad bowl + adding the processed ingredients
- C: carrying the knife + adding the processed ingredients + carrying the processed ingredients
- D: carrying the cutting board + adding the processed ingredients + carrying the knife

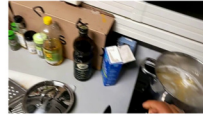
Correct: A

Figure F.10: QID: 9

Q10: What will be my next [NUMBER] actions ?

Rational. Similar to Q9 but using sets of actions performed after the clip.

Design. Same as Q9 but based on the window happening after the current clip.



What will be my next 3 actions ?

- A: carrying the knife + switching the salad bowl + adding the mushrooms
- B: grabbing the pot + carrying the pot
- C: carrying the trivet + stirring the ratatouille + putting the bowl
- D: switching the plate + grabbing the package + switching the bell pepper

Correct: B

Figure F.11: QID: 10

Q11: I am currently [VERB] the [NOUN], what [BE|past,present,future] my [NEXT/CURRENT/PREVIOUS] activity ?

Rational. Prediction of the activity based on the current action.

Design. Similar to Q6,Q7 and Q8 but clips based on actions and answers based on activities.



I am currently grabbing the zucchini, what was my previous activity ?

- A: Cleaning up
- B: Gathering supplies
- C: Setting up and not cooking
- D: Cooking at the stoves

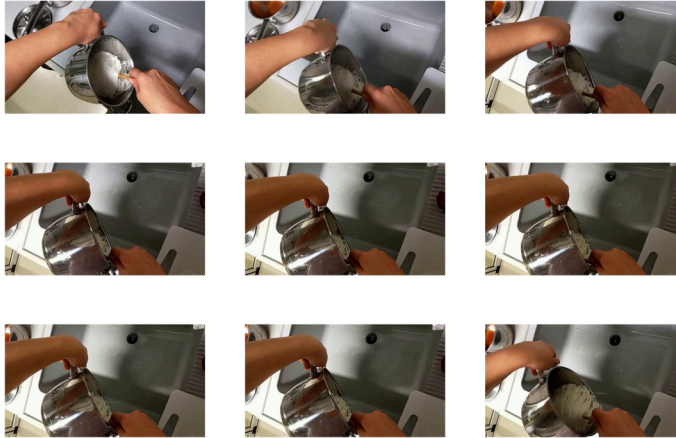
Correct: A

Figure F.12: QID: 11

Q12: I am currently [VERB], what [BE|past,present,future] my [NEXT/CURRENT/PREVIOUS] activity ?

Rational. Prediction of activity based on current action (verb).

Design. Similar to Q11 but using clip based on verbs instead.



I am currently pouring, what is my current activity ?
 A: Preparing ingredients
 B: Gathering supplies
 C: Setting up and not cooking
 D: Cooking at the stoves

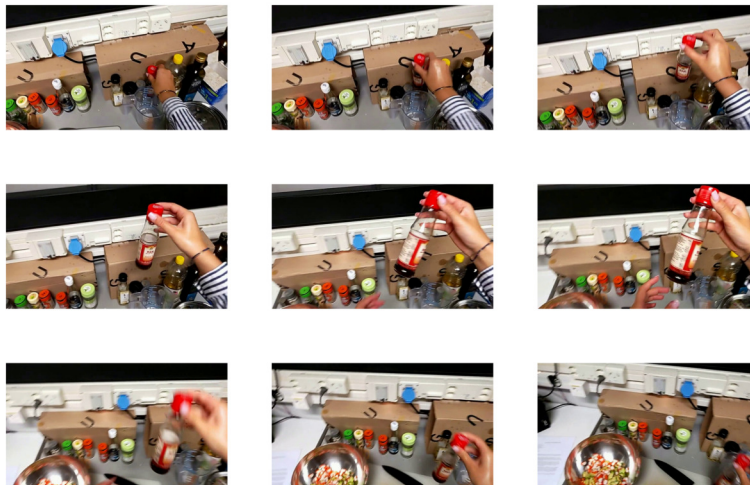
Correct: A

Figure F.13: QID: 12

Q13: What am I [VERB]?

Rational. Prediction of the noun based on the current action (verb).

Design. Similar to Q3, clips are sampled from verb segments, and answers are nouns randomly sampled from the noun list.



What am I carrying ?
 A: The surimi
 B: The tissue
 C: The pot_lid
 D: The bottle

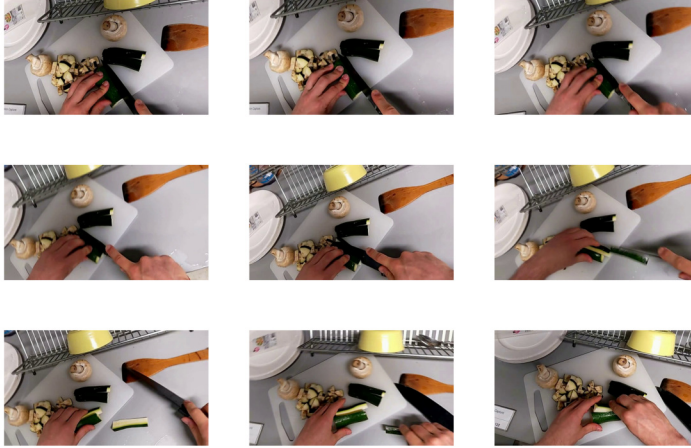
Correct: D

Figure F.14: QID: 13

Q14: How much time passes from [VERB] the [NOUN] to [VERB] the [NOUN] ?

Rational. Estimating timings is difficult for VLMs but also for humans. This question is designed to test the ability of the model to estimate time between two actions. The question is designed to be difficult for humans as well.

Design. The window size is fixed at 200 frames. For each action ending, we look for the previous action starting within the window. The correct answer is the time between the two actions. There is no minimum gap (the answer can be 0). The difficulty is set by the range from which the answers are sampled being 6.67s, 3.33s, and 1.67s around the correct answer for easy, medium, and hard respectively.



How much time passes from cutting the zucchini to grabbing the zucchini?

- A: 0.5s
- B: 1.067s
- C: 1.6s
- D: 2.167s

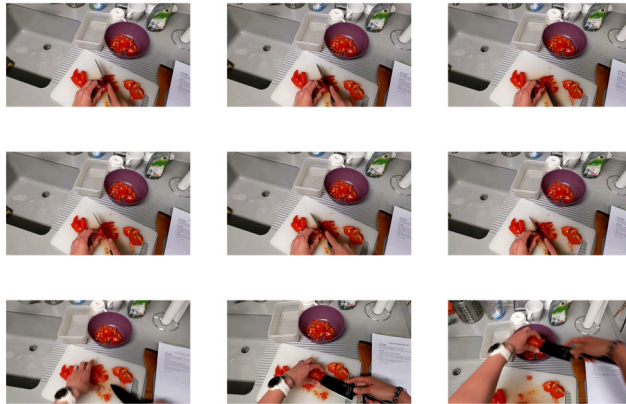
Correct: A

Figure F.15: QID: 14

Q15: At what moment does [VERB] the [NOUN] starts in the clip ?

Rational. Estimating timings is difficult for VLMs but also for humans. This question is designed to test the ability of the model to estimate the time at which an action starts. This relates to part of a traditional action segmentation task. The question is designed to be difficult for humans as well.

Design. The window size is fixed at 200 frames. For each clip, we extract the start of actions within the clip. The correct answer is the time between the start of the action and the start of the clip. The difficulty is set by the range from which the answers are sampled being 6.67s, 3.33s, and 1.67s around the correct answer for easy, medium, and hard, respectively.



At what moment does carrying the tomatoes starts in the clip?

- A: 5.8s
- B: 3.3s
- C: 4.967s
- D: 4.133s

Correct: D

Figure F.16: QID: 15

Q16: What is my age ?

Rational. The intuition is that human behavior changes with age. Using long-term information, can the models precisely estimate the age of the participant? Visual cues can also be present based on the appearance and shapes of the hands, and potentially using the participant's clothing.

Design. The age is a meta information collected during the data collection. The window size is a value in 1000, 5000, or 10000 frames. The difficulty is set by the range from which the answers are sampled, being 30 years, 20 years, and 10 years around the correct answer for easy, medium, and hard respectively.



Figure F.17: QID: 16

Q17: What recipe am I cooking ?

Rational. Participants are cooking one of the following recipes: Omelet (with a tomato salad), Risotto (with an avocado salad), Ratatouille (with pasta), and Pad Thai. While the ingredients can be similar between recipes, the context information is present to infer the recipe even without seeing the outcome of the cooking session.

Design. The recipe is meta information collected during the data collection. The window size is a value in 1000, 5000, or 10000 frames. The answers are sampled from the recipe list.



Figure F.18: QID: 17

Q18: How many times am I [VERB] the [NOUN] in this session ?

Rational. This question is designed to test the ability of the model to count the number of actions in a session. This question can be answered by an estimation based on context and participants' behaviors rather than counting the number of times an action is performed.

Design. The window size is a value in 1000,5000,10000 frames. The correct answer is the number of actions performed in the session. The answers are sampled from a range of 10 around the correct answer.



How many times am I grabbing the peeler in this session ?

- A: 2
- B: 7
- C: 0
- D: 6

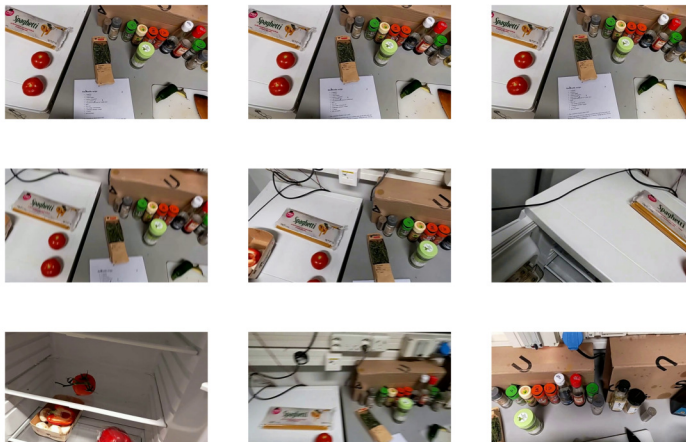
Correct: A

Figure F.19: QID: 18

Q19: How many times am I [VERB] in this session ?

Rational. Same as Q18 but using verbs instead of actions.

Design. Similar to Q18 but using clips based on verbs instead of actions. The range of answers is sampled from a range of 15 around the correct answer, making the question more difficult than Q18.



How many times am I tasting in this session ?

- A: 15
- B: 13
- C: 1
- D: 3

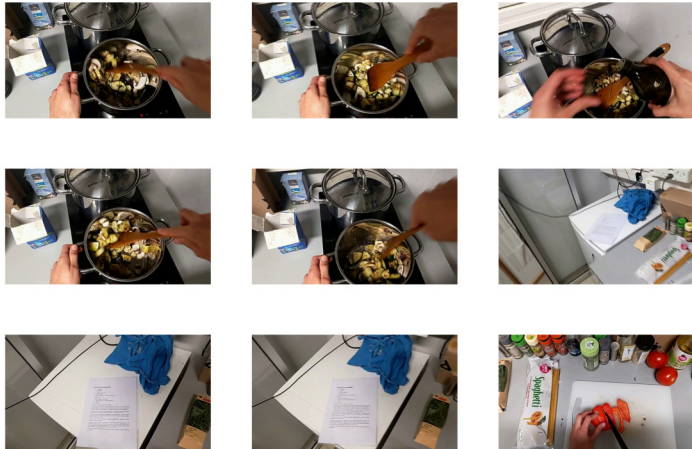
Correct: D

Figure F.20: QID: 19

Q20: For how long am I cooking ?

Rational. This question refers to the total duration of the cooking session. The question is designed to be difficult for humans as well. The model should make an estimation based on the participant's behavior, on the recipe cooked, and the progress along the session.

Design. The window size is a value in 1000, 5000, or 10000 frames. The correct answer is the total duration of the session. The difficulty is set by the range from which the answers are sampled being 30 min, 20 min, and 10 min around the correct answer for easy, medium, hard respectively.



For how long am I cooking ?

- A: 37min
- B: 32min
- C: 28min
- D: 39min

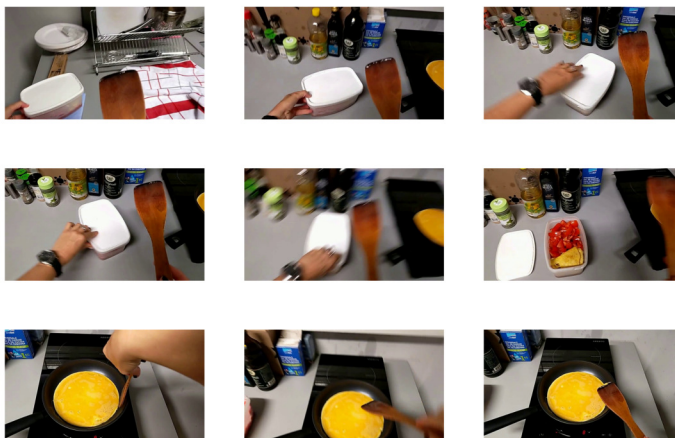
Correct: B

Figure F.21: QID: 20

Q21: What is the correct sequence of actions?

Rational. Ordering actions requires the model to understand the context of the actions and their temporal order. This question is designed to be difficult for humans as well.

Design. The window size is set to 500 frames. We select 3,4,5 actions from the clip (easy, medium, hard). Answers are the selected actions in permutated orders. Note that the same action can be present multiple times in the sequence.



What is the correct sequence of action ?

- A: opening the box + grabbing the pan + holding the pan + putting the box + stirring the omelet
- B: putting the box + moving the box + opening the box + grabbing the pan + holding the pan + stirring the omelet
- C: stirring the omelet + grabbing the pan + moving the box + putting the box + holding the pan
- D: grabbing the pan + putting the box + holding the pan + opening the box + stirring the omelet

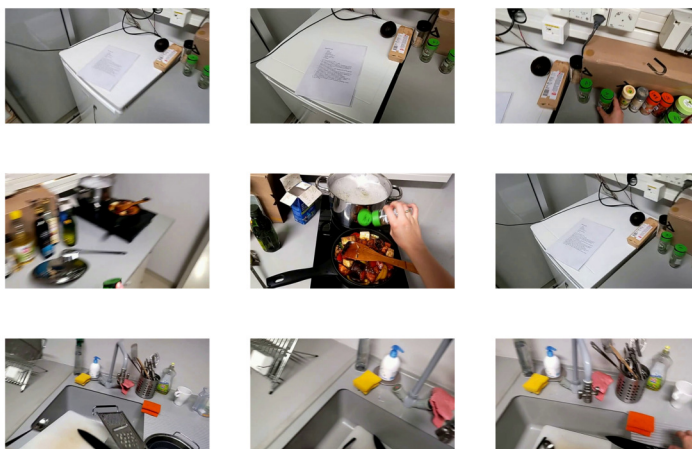
Correct: B

Figure F.22: QID: 21

Q22: What is the longest action in this session ?

Rational. This question asked about the longest action from the propositions in the whole session. This requires the model to have an estimation of how long actions are and add them to cues from the recipe context and participant's behavior. This question is designed to be difficult for humans as well.

Design. We compute the duration of all actions in the session and sort them. We then sample from the whole list of durations, from a window of 9 subsequent action durations, or from a window of 4 subsequent action durations (easy, medium, hard). The correct answer is the longest action, and other answers are sampled randomly from these windows.



What was the longest action in this session ?

- A: Open
- B: Close
- C: Read
- D: Cut

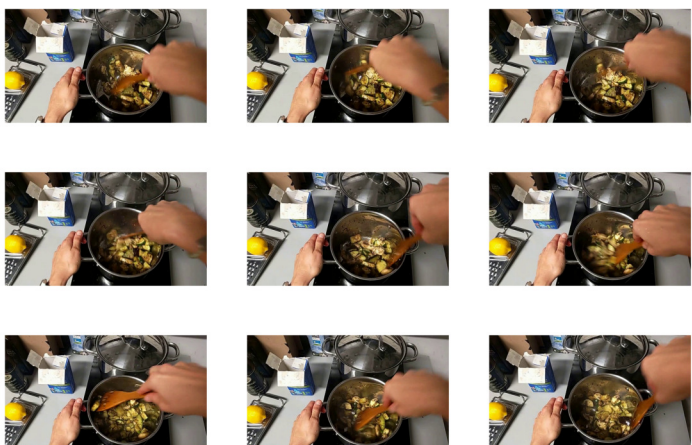
Correct: C

Figure F.23: QID: 22

Q23: What is the average height of my eyes in this clip ?

Rational. This question is designed to test the ability of the model to estimate the height of the participant's eyes. The model should extract context from the environment and the egocentric view of the participant; the answer should also be a reasonable value.

Design. The window size is set to 50 frames. The correct answer is extracted from the pose estimation of the participant (average of eye keypoints). The difficulty is set by the range from which the answers are sampled being 90%, 50%, and 20% of the correct answer around the correct answer for easy, medium, and hard, respectively.



What is the average height of my eyes in this clip ?

- A: 1.542m
- B: 1.322m
- C: 1.763m
- D: 0.881m

Correct: B

Figure F.24: QID: 23

Q24: What is the average shape of my [RIGHT/LEFT] hand in this clip ?

Rational. Due to occlusion in monocular videos, inferring the shape of the hands can be rather difficult. In this question, we leverage the pose estimation data to categorize the hand into four different categories: open, closed, pointed, and pinched. The question can be answered by direct observation of the hands, but also based on the context (e.g., if the participant is carrying a knife).

Design. Hand shapes are extracted from the hand pose estimation data. First, the distance between the fingertips and the wrist keypoint is used to define open vs. closed hand. We then compute the distance between the index finger and the average of the other fingers to define pointed. And finally, we compute the distance between the index and the thumb tip to define pinched. The answer propositions are each possible hand shape option. Clips correspond to frames on which the hand was detected by the HoloLens 2 device. Therefore, the window size can be very short.



What is the average shape of my left hand in this clip ?

- A: pointing
- B: open
- C: closed
- D: pinching

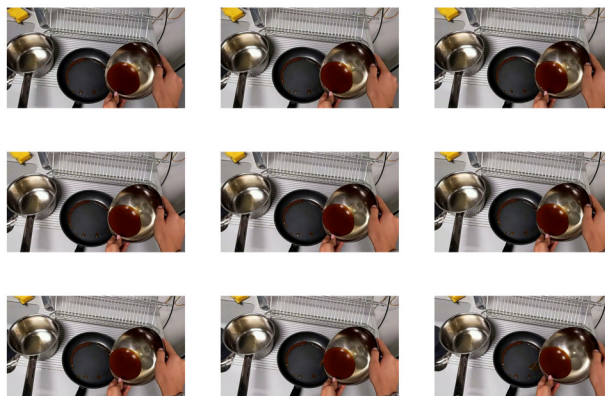
Correct: B

Figure F.25: QID: 24

Q25: What is my average bending angle in this clip ?

Rational. Bending angle can be related to multiple factors: the participant's age, the behavior, or the environment. This question is designed to test the ability of the model to estimate the bending angle of the participant. The model should extract context from the environment and the egocentric view of the participant; the answer should also be a reasonable value (e.g. it cannot be more than 180 degrees).

Design. The window size is set to 50 frames. The correct answer is extracted from the pose estimation of the participant (average of elbow keypoints). The difficulty is set by the range from which the answers are sampled being 100 deg, 50 deg, and 20 deg around the correct answer for easy, medium, and hard, respectively.



What is my average trunk bending angle in the clip?

- A: 144 deg
- B: 172 deg
- C: 199 deg
- D: 166 deg

Correct: B

Figure F.26: QID: 25

Q26: The clip lasts [NUMBER] seconds, what is the average speed of my [RIGHT/LEFT] hand in this clip ?

Rational. Estimation of speed is difficult both for humans and for VLMs. This question requires an estimation of the total distance traveled by the hand in the clip and to divide it by the clip duration. The model should extract context from the environment and the egocentric view of the participant.

Design. The clips are sampled from the verb segments. The correct answer is extracted as the speed of the wrist of the target hand. The difficulty is set by the range from which the answers are sampled being 1 m/s, 0.5 m/s, and 0.1 m/s around the correct answer for easy, medium, and hard, respectively.

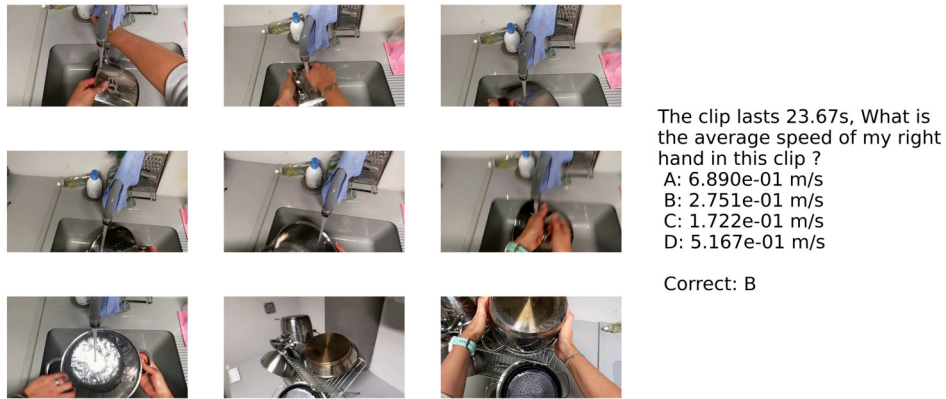


Figure F.27: QID: 26

Q27: The clip lasts [NUMBER] seconds, at what speed am I reaching for the [NOUN] in this clip ?

Rational. Similar to Q26 but specifically for reaching actions on a target object. This gives the model a more precise context to estimate the speed.

Design. The clips are sampled from action segments with "reach" as the verb. The correct answer is extracted as the speed of the wrist of the target hand. The difficulty is set by the range from which the answers are sampled being 1 m/s, 0.5 m/s, and 0.1 m/s around the correct answer for easy, medium, and hard respectively. We filter clips in which no hand is detected by the HoloLens 2 device.

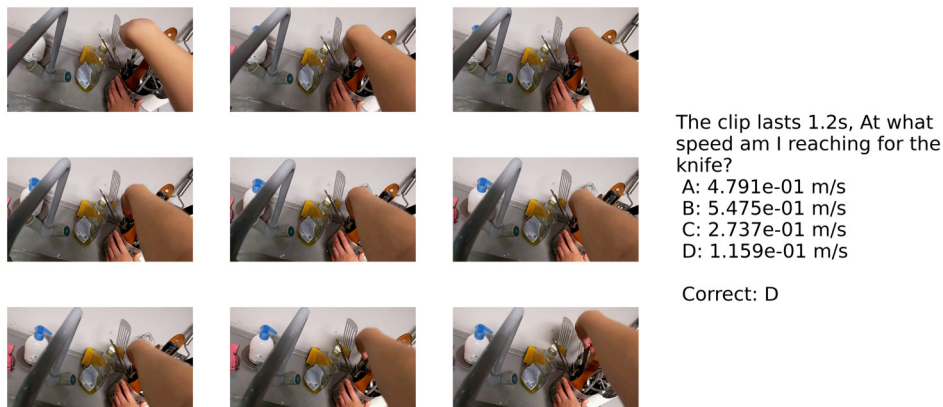
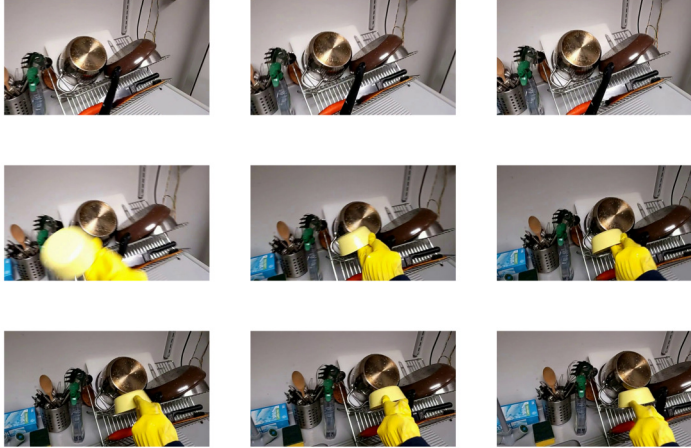


Figure F.28: QID: 27

Q28: The clip lasts [NUMBER] seconds, at what speed am I putting the [NOUN] in this clip ?

Rational. Similar to Q27 but targeted around the putting action

Design. Similar to Q27 but using clips based on the "put" verb.



The clip lasts 1.6s, At what speed am I putting the bowl down?

- A: 0.000e+00 m/s
- B: 6.849e-02 m/s
- C: 1.416e-01 m/s
- D: 2.123e-01 m/s

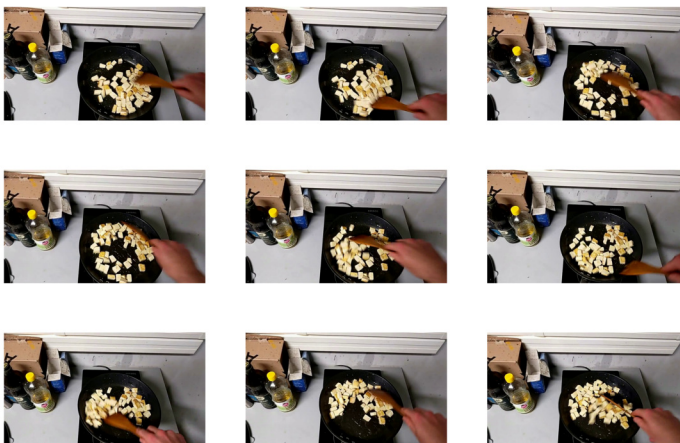
Correct: B

Figure F.29: QID: 28

Q29: What is the [minimum/average/maximum/range of] angle(s) of my [LEFT/RIGHT][BODYPART], of my [LEFT/RIGHT][BODYPART], and my [LEFT/RIGHT][BODYPART] in this clip ?

Rational. Angles are very relevant in behavior analysis. This question challenges VLMs to infer angles from context. This question is designed to be difficult for humans as well.

Design. The window size is set to 50 frames. Possible questions are, for each size, the angle between the elbow, the shoulder, and the hip; the angle between the shoulder, the elbow, and the wrist; the angle between the hip, the knee, and the ankle; and the angle between one shoulder, the other shoulder, and the related elbow. Questions can be about the average, maximum, minimum of the range of the angles present in the clip. The difficulty is set by the range from which the answers are sampled being 100 deg, 50 deg, and 20 deg around the correct answer for easy, medium, and hard respectively.



What is the minimum angle between my left shoulder, my right shoulder and my right elbow in this clip ?

- A: 119.5 deg
- B: 97.3 deg
- C: 102.9 deg
- D: 130.6 deg

Correct: C

Figure F.30: QID: 29

Q30: What is the [minimum/average/maximum] distance between my hands ?

Rational. This question aims to test the ability of models to estimate distances based on context. This question is designed to be difficult for humans as well.

Design. The clips are extracted from frames in which both hands are detected by the HoloLens 2 device. The correct answer is extracted from the distance between the two wrists. The difficulty is set by the range from which the answers are sampled being 1m, 0.5m and 0.2m around the correct answer for easy, medium, and hard respectively.

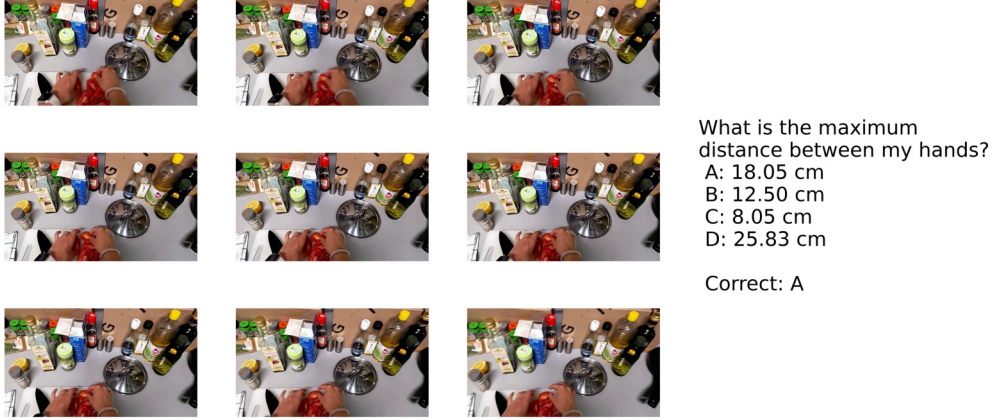


Figure F.31: QID: 30

F.2 Visual Language Model (VLM) Evaluation

We evaluate recent state-of-the-art VLMs, including InternVL2.5-8B [11], LLaVA-OneVision-7B [36], Qwen2.5-VL [5], and Gemini 2.0 Flash [15], using the `lmms-eval` framework [100]. For Qwen2.5-VL, we report results for both the 7B and 32B parameter variants to examine the impact of model scale.

Dataset and Preprocessing.

We uniformly sample N frames from each video segment to serve as the model’s visual context. Unless otherwise specified, we used $N = 8$ frames per query. For Qwen2.5-VL 32B, we set $N = 4$ due to computational constraints. The frames are sampled evenly between the annotated start and end timestamps of each clip.

Prompt Construction. For each question, the input prompt presented to the model follows a standardized template:

Prompt Used for LEMONADE QA Benchmark
<p>Answer the following multiple-choice question using the given images. Question: [question text] Choices: A. [option 1] B. [option 2] C. [option 3] D. [option 4] Respond only with the letter of the correct answer.</p>

For each question, the models generate up to 128 output tokens using greedy decoding ($T = 0$). Models are explicitly instructed to respond with a single letter corresponding to their predicted answer. We parse responses using rule-based heuristics to extract the answer choice from each response. In cases where no valid answer can be identified, we default the prediction to “A” (which is random, as the order in the question is randomized).

F.3 Egocentric reprojection of the pose data









All poses are projected to their egocentric reference frame before being used in the action recognition, the action segmentation and the full-body motion generation benchmarks. We will describe the projection method in this section. The reference frame and reprojection matrix are calculated using the body joints as reference. After extracting the centroid, we use the average between joints on the left and right sides of the body to get a primary reference vector. We then extract a temporary vector by taking the average between the lower and upper body parts. By taking the cross product of both vectors, we can extract a second reference vector. Finally, by taking the cross product between the two reference vectors, we obtain the third orthogonal reference vector. The hand poses and eye gaze are also reprojected using the reference frame calculated in the previous step. This can imply implicit arm or head positions when using this information.













F.4 Details on the Action Recognition Benchmark

Data from different modalities are first sent to different modality-specific projection layers to obtain the query tokens and then sent to the transformer backbone to extract cross-modality token-wise features. Finally, features from different modalities are merged using Average Pooling to make action predictions. Specifically, for video-type data, we follow the original implementation of VideoMAE [83], select 16 frames and split into 16×16 patches for each frame; we introduce the egocentric view and one global exocentric view as inputs in our baseline designs. For point-type data (e.g., hand-body poses and eye gaze), we sample poses of 32 frames. For every pose, we concatenate all the pose joints and their corresponding confidences together and apply a linear projection layer to generate a query token. We tried two different variants that either use VideoMAE [83] weights pre-trained on the EPIC-KITCHENS-100 [14] or trained from scratch. We follow standard practice [7, 23] to predict verb class and noun class separately and then apply the outer product to obtain the action class. The code to reproduce the results is available at <https://github.com/amathislab/EPFL-Smart-Kitchen>

F.4.1 Additional results

Table F.3 provides a comparison of the VideoMAE models trained from scratch using different input modalities on the action recognition benchmark. It is important to note that these results are significantly lower compared to those from models pretrained on EPIC-KITCHENS-100.

Table F.3: **Fine-grained action recognition benchmark results trained from scratch.** : ego-centric view, : global exocentric view, : 3D body pose, : 3D hand pose, : eye gaze,   \times multiple  : hand cropped videos. Combining modalities has the potential to increase the performance. Our best results are achieved by cleverly merging these modalities together.

Modalities	All Classes Accuracy Top1/5			Head Classes Accuracy Top1/5			Tail Classes Accuracy Top1/5		
	Action	Verb	Noun	Action	Verb	Noun	Action	Verb	Noun
	16.76/35.29	41.85/86.65	29.35/59.92	19.40/40.60	44.86/89.17	33.12/65.81	1.51/4.64	24.51/72.13	7.58/25.94
	14.29/31.32	42.89/84.77	23.51/48.95	16.58/35.69	44.98/87.12	26.60/54.03	1.08/6.11	30.81/71.20	5.68/19.64
	16.97/35.75	42.35/87.06	29.46/61.20	19.77/41.22	45.66/89.50	33.30/67.59	0.81/4.17	23.23/72.98	7.31/24.35
	13.27/27.32	40.63/76.55	22.01/43.12	15.30/30.79	42.07/78.34	24.87/46.99	1.58/7.31	32.35/66.22	5.49/20.76
	17.73/35.85	48.81/87.29	26.74/52.50	20.42/40.48	50.55/89.09	30.15/57.51	2.20/9.14	38.77/76.85	7.07/23.64
 	16.90/35.12	47.60/86.36	25.93/52.00	19.48/39.52	49.17/88.22	29.35/56.89	2.01/9.70	38.58/75.61	6.18/23.81
 	17.01/34.81	48.09/86.92	26.03/51.33	19.63/39.14	49.74/88.79	29.47/55.95	1.93/9.86	38.54/76.11	6.15/24.70
  \times multiple 	21.37/44.37	48.52/90.22	35.73/68.36	24.52/50.17	51.25/92.50	39.88/73.96	3.21/10.90	32.74/77.04	11.79/36.03

F.4.2 Data Preprocessing

We adapt VideoMAE [83] model with the VIT-L backbone [16] to build our multimodal baselines. Data from different modalities are projected into query tokens, processed by a transformer for feature extraction, and merged via Average Pooling for action prediction. Video data uses VideoMAE [83] with 16 frames, incorporating egocentric and global exocentric views. Point-type data (e.g., hand-body poses, eye gaze) samples 32-frame poses, concatenates joint positions with confidence scores, and linearly projects into query tokens. We test VideoMAE [83] weights pre-trained on the EPIC-KITCHENS-100 [14] or trained from scratch. Action classes are derived by separately predicting verbs and nouns, then combining them via outer product [7, 23].

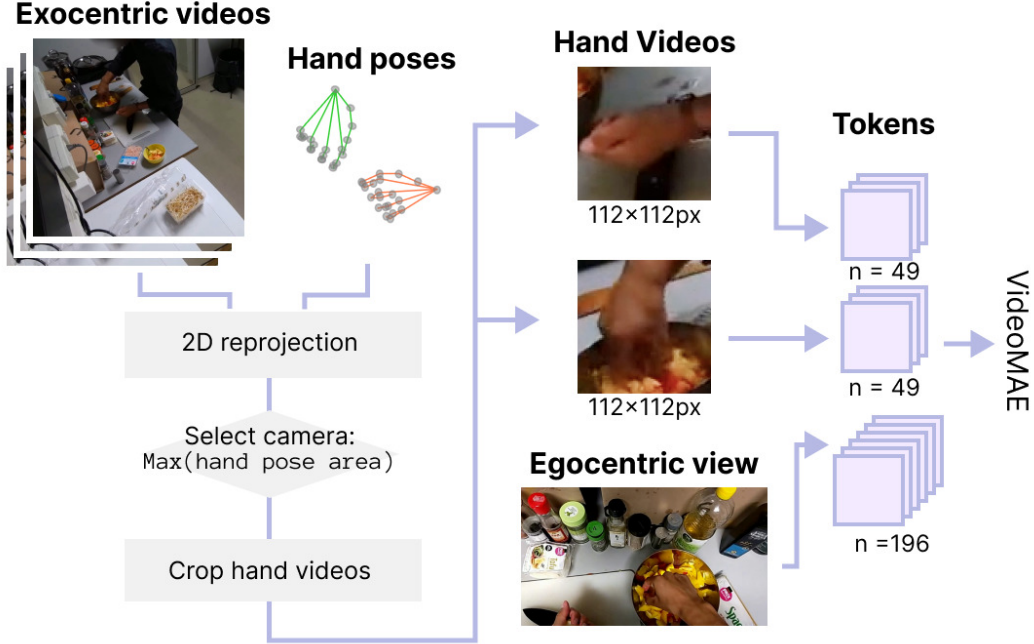


Figure F.32: Feature extraction from multimodal data. Hand videos are extracted from the exocentric view which observes the largest hand area and used together with the egocentric to VideoMAE.

As is discussed in the main paper, our EPFL-Smart-Kitchen-30 has videos (35.9 ± 11.6 -minute-long) that are longer than the existing action datasets. To achieve efficient data loading, we used the framework from AVION [103] to pre-resize the original videos and cut them into 30-second-long chunks. For a certain action annotation queried in the annotation list, we only load the chunks that cover the annotated segment.

F.4.3 Hyperparameters

We adapted a condensed version of VideoMAE [83] from TIM [7], which provides the pre-trained weights on the EPIC-KITCHENS-100 dataset [14] and modifies the original classification head to two classification heads so that the model can predict verb and noun classes separately. We use the similar hyperparameters as TIM [7] to train the VideoMAE [83] model. Specifically, for an input action segment, we evenly sample 16 frames for the video inputs (egocentric view and exocentric view) and sample 32 frames for the pose inputs (eye, hand, and body poses). The initial learning rate is set to 0.0003 and is controlled by the cosine scheduler. We additionally set weight decay as 0.05 and drop-out rate as 0.3 to prevent overfitting.

F.4.4 Hand-cropped video extraction

We implemented a method to merge the exocentric view and the hand pose data in a slightly more sophisticated manner. Specifically, for a given action clip, we project the 3D hand poses onto all camera views, select the view displaying the largest hand region, and crop the frames around both hands. We send these hand-cropped videos together with the egocentric view video as input to the VideoMAE model, splitting the cropped video into 8×8 patches for each hand (see Figure F.32)

F.5 Details on the Action Segmentation Benchmark

F.5.1 Models

DLC2Action [32] adapted image-based models to process pose estimation data. MS-TCN3 is a modified version of MS-TCN++ [39]. The main difference between MS-TCN3 and MS-TCN++ is combining the output of the last and second-to-last layers of the first stage to be passed as input to

the second stage, in order to allow for rich representations. C2F-Transformer is a modification of C2F-TCN [75] that replaces some convolution operations with attention. The code to reproduce the results is available at <https://github.com/amathislal/EPFL-Smart-Kitchen>

F.5.2 Additional results

Table F.4 presents the comprehensive results obtained from the action segmentation benchmark, featuring the Edit Score and segmental F1 measured at a 50% threshold.

F.6 Fine-grained performance in action segmentation body vs hands

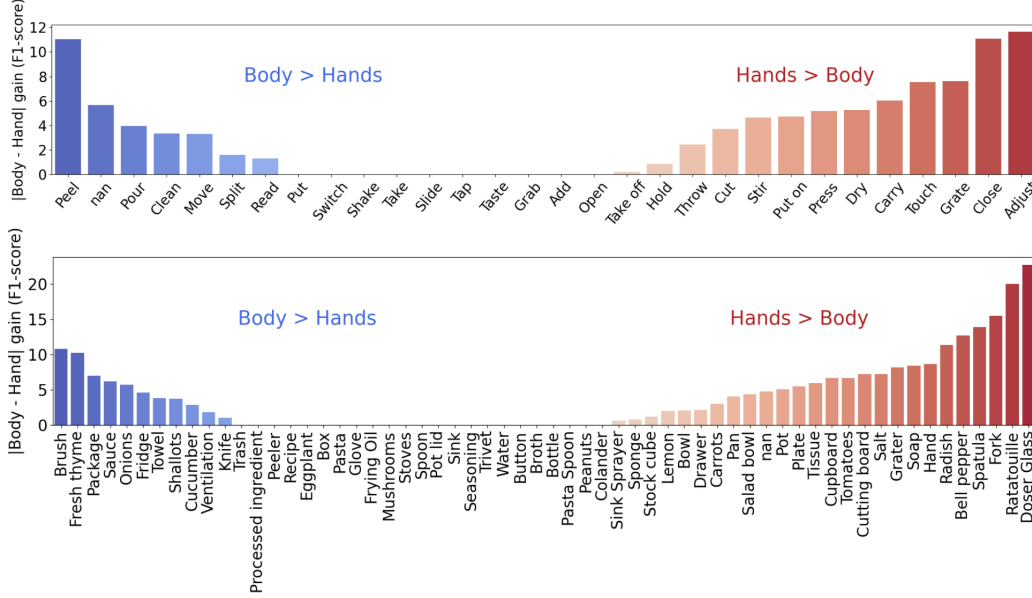


Figure F.33: Absolute difference for action-wise performances using body pose vs hand pose for verbs (top) and nouns (bottom).

Table F.4: **Results for action segmentation benchmark.** 🧑: 3D body pose, 🧤: 3D hand pose, 👁: eye gaze, 📹: egocentric view. * indicates models modified to integrate pose as input data instead of deep image features.

	F1-score ↑								Edit Score ↑								F1@50 ↑							
	🧑	🧤	👁	🧑🧤	🧑👁	🧤👁	🧑🧤👁	📹	🧑	🧤	👁	🧑🧤	🧑👁	🧤👁	🧑🧤👁	📹	🧑	🧤	👁	🧑🧤	🧑👁	🧤👁	🧑🧤👁	📹
Verbs																								
MS-TCN3	18.1	20.2	11.7	20.9	21.1	30.1	33.9		83	84	60	87	87	91	93		2.9	3.9	2.0	4.4	4.3	5.5	10.4	
C2F-TCN*	18.8	20.1	12.2	22.1	22.2	34.6	36.6		86	86	72	89	88	93	94		7.6	7.4	3.3	8.5	9.3	16.7	19.0	
C2F-Transf.	19.9	22.4	13.1	22.8	22.2	35.0	39.6		87	88	73	88	88	93	95		6.2	7.3	3.1	8.5	9.3	16.6	19.1	
EDTCN*	19.6	23.0	11.9	22.1	23.2	34.3	39.6		87	88	66	88	90	93	95		11.5	13.4	5.8	12.8	14.9	23.0	27.4	
Nouns																								
MS-TCN3	10.6	13.4	7.6	15.6	11.3	31.2	35.9		85	87	69	84	86	95	95		2.4	2.5	1.3	3.0	4.2	5.0	9.2	
C2F-TCN*	12.0	14.3	7.9	16.1	10.8	35.2	41.3		89	89	81	90	80	96	96		4.6	5.8	2.5	6.0	2.7	15.8	19.3	
C2F-Transf.	11.1	12.9	7.8	13.4	9.2	29.0	37.2		85	85	73	86	72	94	96		3.1	3.9	1.9	4.0	4.1	7.8	11.2	
EDTCN*	11.9	11.2	7.1	12.3	11.9	24.3	37.3		82	83	65	83	81	92	95		5.3	4.5	2.4	5.4	1.5	11.8	18.4	
Activity																								
MS-TCN3	51.8	58.6	31.9	54.4	58.5	72.9	66.4		88	90	85	89	90	94	92		15.7	13.3	8.7	13.9	15.1	24.6	37.9	
C2F-TCN*	54.5	55.4	41.3	61.8	61.2	72.2	65.5		87	87	81	89	89	93	90		26.3	23.6	13.8	28.7	28.6	46.4	46.3	
C2F-Transf.	51.2	56.9	38.8	62.1	59.9	70.5	67.7		88	89	83	90	90	93	92		27.3	28.3	18.3	37.0	35.2	48.3	49.2	
EDTCN*	49.0	53.5	32.0	53.1	54.2	71.0	67.4		86	88	85	88	88	93	92		22.5	24.7	12.6	27.0	26.4	50.0	44.5	

F.6.1 Data Preprocessing

To deal with long durations of videos, we sequence the videos into subsequences of constant lengths (see segment lengths in Table F.5) with an overlap of 10%. These segments are sampled at a fixed ratio defined as temporal subsampling in Table F.5.

F.6.2 Hyperparameters

In the action segmentation benchmark, we searched for the best hyperparameters using Optuna [2] for both nouns and verbs using Exo-Body as input. We used the same parameters when training for other modalities. All models were trained with a batch size of 512. The best parameters are outlined for each model in Table F.5

Table F.5: Hyperparameter used for the action segmentation benchmark.

HyperParameters	MS-TCN3	C2F-TCN	C2F-Transformer	EDTCN
Nouns				
Loss - alpha	$5.14e^{-5}$	$1.4e^{-4}$	$2.0e^{-5}$	$3.4e^{-5}$
Loss - focal	✗	✗	✓	✗
Temporal subsampling	0.84	0.95	0.90	0.90
Number f maps	106	59	32	
Segment length	256	512	512	256
Learning rate	$1e^{-5}$	$1e^{-5}$	$1e^{-5}$	$1e^{-5}$
Batch size	512	512	512	512
Verbs				
Loss - alpha	$5.69e^{-4}$	$1.77e^{-3}$	$4.42e^{-5}$	$5.66e^{-3}$
Loss - focal	✗	✗	✓	✗
Temporal subsampling	0.91	0.83	0.85	0.95
Number f maps	118	58	64	
Segment length	512	1024	1024	128
Learning rate	$1e^{-5}$	$1e^{-5}$	$1e^{-5}$	$1e^{-5}$
Batch size	512	512	512	512
Activity				
Loss - alpha	$5.99e^{-5}$	$9.21e^{-4}$	$1.67e^{-5}$	$2.74e^{-3}$
Loss - focal	✗	✗	✓	✓
Temporal subsampling	0.82	0.75	0.94	0.80
Number f maps	123	128	128	
Segment length	256	1024	512	256
Learning rate	$1e^{-5}$	$1e^{-5}$	$1e^{-5}$	$1e^{-5}$
Batch size	512	512	512	512

Additionally, the MS-TCN3 model was trained with 14 PG layers and 10 R layers.

F.6.3 Feature engineering details

We list below the calculations performed for each feature extraction.

speed, acceleration, speed direction, joint angles, joint acceleration, intra-coordinate distances, distances to centroid, and centroid position

- *Joint speed/acceleration* : speed and acceleration from the raw coordinates.
- *Joint angles*: After projection of paired vectors in their respective 2D planes.
- *Speed direction*: 3D vector taken as the difference between joint speed at time $t + 1$ and at time t .
- *Intra-coordinate distances*: Pairwise distance between raw coordinates.
- *Centroid*: Mean value of raw coordinates at a given frame.
- *Distance to centroid* : Distance between each coordinate and the centroid.

Holo-Hands vs Exo-Hands. Using exocentric information to infer hand pose information is not only more accurate, but also significantly increases the performance for all action segmentation models (Table F.6).

Table F.6: Performance comparison between HoloLens pose estimation and Exocentric pose estimation. indicates models that have been modified to integrate pose as input data instead of deep image features.

	Verbs		Nouns	
	Holo-👤	Exo-👤	Holo-👤	Exo-👤
MS-TCN3*	14.0	20.2	11.3	15.6
C2F-TCN*	13.6	20.1	9.7	14.5
C2F-Transformer*	15.8	22.4	20.2	14.0
EDTCN*	15.3	23.0	10.6	14.6

F.7 Details on the Full-body Motion Generation Benchmark

For the full-body motion representation, we use redundant information to serve as input motion representation, which can usually get a more robust representation. Specifically, we concatenate both the joint locations (body, hands and eye gaze) and the joint angles (body, hands and eye gaze) as the motion representation. We obtain a 327-dim motion representation. We use our fine-grained actions as the input text, coupled with tags (e.g., *VERB*, *NOUN*, *ADV* for adverbs) for each word. They are then transformed into word tokens using CLIP’s text encoder [66]. We similarly process the middle egocentric view frame into visual features using CLIP’s image encoder [66]. As baselines, we adapted two recent motion generation models, i.e., T2M-GPT [99] and MoMask [27]. MoMask is the state-of-the-art model on HumanML3D [26] and T2M-GPT is another strong model working with a different mechanism. We trained models with two text prompts: verb-noun pairs (actions) and verbs only. The code to reproduce the results is available at <https://github.com/amathislab/EPFL-Smart-Kitchen>

F.7.1 Data Preprocessing

We follow the HumanML3D [26] codebase to preprocess the EPFL-Smart-Kitchen-30 pose data. The input to the motion generation model is a processed pose feature of size 263, which includes the pose joint positions, joint rotations, as well as ground contact information. To make the training phase of the motion generation model stable, we only keep the segments with frame numbers larger than 64 and smaller than 300 for training, aligning with the approach used in HumanML3D [26].

F.7.2 Evaluator training

Both the FID and R precision require a pretrained model to extract features first and then compute the metrics. Therefore, we also re-train the evaluator on our dataset following the HumanML3D [26] codebase.

Table F.7: Table comparing different models and text types on rFID and MPJPE metrics.

Text type	Verbs		Actions	
	rFID ↓	MPJPE ↓	rFID ↓	MPJPE ↓
VQ	1.819	0.295	1.568	0.301
R-VQ	0.606	0.292	0.732	0.292

F.7.3 Tokenizer performance

T2M-GPT [99] used the vector quantization (VQ) [87] to serve as the tokenizer while MoMask [27] proposed to use residual vector quantization (R-VQ) as a tokenizer. We train both of them on our dataset so that T2M-GPT [99] and MoMask [27] can use them as model components. Therefore, we also show the reconstruction performances of VQ and R-VQ (Table F.7), which are measured by the reconstruction Fréchet Inception Distance (rFID) and the Mean Per Joint Position Error (MPJPE).

G Datasheets for datasets

We follow the template provided by Gebru et al. [22] to supplement the dataset to document the motivation, composition, collection process, recommended uses, distribution, and maintenance. For clarity, we select and show only the relevant sections from the template. Notice that all sections previously described in the supplementary material will be simply referred to by their respective sections. Additionally, all questions that give clues on the identity of the group will be omitted from present restricted answers.

G.1 Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created to enable research on human motor function and fine-grained motor behaviors. The dataset was created intentionally to motivate the community in the development of new behavior understanding methods and with the prospect of analyzing the behavior of patients with motor disorders such as stroke and amputee patients.

G.2 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description

The dataset consists of RGB videos, depth, IMU sensors, and eye gaze of subjects cooking in a kitchen together with action annotations. There is only one subject cooking at the time, but there might be very short interactions with the experimentalists for clarifications.

How many instances are there in total (of each type, if appropriate)?

There are 49 recorded sessions of each type.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

This dataset contains all currently processed instances but will later be augmented with additional young healthy, but also older, participants as well as neurological patients. The current sample is representative of the young healthy cohorts, but this has not been validated because of the lack of available methods of comparison. (we do not expect consistency in cooking behaviors in the young healthy cohort).

What data does each instance consist of? “Raw” data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance consists of a set of ten RGB videos, nine depth videos (all static cameras), eight IMUs with 3-channel recordings, and action annotations (33 verbs + 79 nouns + 6 activities) of cooking behaviors.

Is there a label or target associated with each instance? If so, please provide a description.

Each instance is labeled with the participant’s encoded identity and the session number given as a date.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No information is missing, all mentioned data is included.

Are relationships between individual instances made explicit (e.g., users’ movie ratings, social network links

There is no relationship between individual instances. Participants might be related, but this should have no direct impact on their body motions.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them. Data splits (training/validation/testing) are provided with the dataset and are used to train the baselines mentioned in the main paper.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description. There are possible errors in the 3D pose estimation when the body or hands are visible from too few cameras (extreme positions in the kitchens), but this should only have a few overlaps with annotated actions.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' nonpublic communications)? If so, please provide a description.

All videos are anonymized and not confidential. All subjects have consented to the public sharing of their recordings.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

The current version of the dataset does not identify any subpopulations. Genders mentioned in the main paper were given according to the subjects' self-identifications.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

All videos have been anonymized; it is not possible to identify the individuals present in the videos.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No.

G.3 Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The video data and other recordings are directly observable from the data. The action annotations are indirectly inferred from the videos by watching the subject's movements. Details are given in the main paper.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?

The hardware details used for the data collection are given in section A.1. We used Microsoft SDK and Developer tools to process the Kinect Azure and HoloLens 2 data. We additionally wrote Python scripts to collect and organize the data collection. The resulting data (videos) were validated with the qualitative visual investigation for each individual session during data collection and during data preprocessing.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

N/A

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Subjects participating in the study were voluntary participants from external advertisement. They were compensated with the equivalent of 22.5\$ per hour spent in the kitchen and were allowed to eat the resulting dish prepared by themselves. Experimentalists were eventually compensated with undesired meals.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The dataset was collected over the course of a year.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation

The ethical protocol was reviewed by an IRB. We can currently not provide documentation related to the matter to preserve anonymity.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

We collected the data from the individuals in question directly.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

The individuals in question were notified about the data collection; for further details, see section D.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

Yes, see D

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

Yes, see D

Any other comments?

None.

G.4 Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section

Preprocessing steps consist of synchronization (truncation of video data to start and stop of experiments) (see section A.2) and 3D pose extraction (see section C). Additionally, we anonymized the

data by blurring all faces present in videos. For the action annotations, we filter certain actions in the benchmarks (detailed in the main paper).

Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the “raw” data.

The raw data (untruncated unanonymized videos) are not shared in the dataset.

Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.

The code will be shared upon acceptance of the submission.

G.5 Uses

Has the dataset been used for any tasks already? If so, please provide a description.

The dataset has not been used yet.

What (other) tasks could the dataset be used for?

The dataset could be used in language-associated tasks using the verb description, for motion analysis using the 3D pose data, different type of multi-feed integration to predict one or the other modality, and in many more applications.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

Anonymization locally harms the quality of a video, but it should not have an impact on future usage in mentioned tasks;

Are there tasks for which the dataset should not be used? If so, please provide a description

This dataset was collected solely in the behavior analysis domain. This dataset should not be used for unrelated tasks.

G.6 Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The dataset is publicly available on Zenodo. For more details, please refer to Table 6.

How will the dataset be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset is distributed on Zenodo (10.5281/zenodo.15535461). The code to reproduce the results is also available on Github : <https://github.com/amathislab/EPFL-Smart-Kitchen>.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

Yes, the dataset is distributed under a CC-BY 4.0 license.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No other party imposed IP-based or other restrictions on the data.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

None.

G.7 Maintenance

Who will be supporting/hosting/maintaining the dataset?

The authors will personally support, host and maintain the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Email addresses are provided on the referred website. Issues can also be raised on the GitHub page.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?

If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)?

The dataset will be updated with additional instances and eventual corrections. All changes will be stipulated on the website hosting the dataset.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question informed that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

The applicable limits are not defined in terms of time; rather, they apply whenever the anonymization was conducted and the dataset is public. The raw data (untruncated and unanonymized) is to be deleted upon publication.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.

Older versions of the dataset will be maintained, hosted, and supported since the newer versions will only be an enlargement of the data size with new cohort groups and, therefore, are still dependent on older versions of the dataset.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description. Prior contact with the authors must be conducted before any extension or augmentation of the dataset. The authors need to validate the application, usage, and accuracy of the modifications eventually made to the dataset.